UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA
DE TRANSPORTES

MARIA LÍGIA CHUERUBIM

**Possibilidades em Inteligência Artificial na detecção de
padrões e previsão de acidentes em rodovias**

São Carlos
2019

MARIA LÍGIA CHUERUBIM

**Possibilidades em Inteligência Artificial na detecção de
padrões e previsão de acidentes em rodovias**

**Programa de Pós-graduação em Engenharia
de Transportes da EESC-USP**

**Exemplar definitivo (corrigido). O exemplar
original está disponível na CPG da EESC-USP**

São Carlos, **26/04/2019**

RESOLUÇÃO CoPGr N º 6018, DE 13 DE OUTUBRO DE 2011, artigo 5º

Tese apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo, para a obtenção do título de Doutor em Ciências - Programa de Pós-Graduação em Engenharia de Transportes.

Área de concentração: Infraestrutura de Transportes - Geomática.

Orientador: Prof. Dr. Irineu da Silva

São Carlos

2019

AUTORIZO A REPRODUÇÃO TOTAL OU PARCIAL DESTE TRABALHO,
POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS
DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da
EESC/USP com os dados inseridos pelo(a) autor(a).

C559p Chuerubim, Maria Ligia
Possibilidades em Inteligência Artificial na
detecção de padrões de e previsão de acidentes em
rodovias / Maria Ligia Chuerubim; orientador Irineu da
Silva. São Carlos, 2019.

Tese (Doutorado) - Programa de Pós-Graduação em
Engenharia de Transportes e Área de Concentração em
Infraestrutura de Transportes -- Escola de Engenharia
de São Carlos da Universidade de São Paulo, 2019.

1. Banco de dados. 2. Acidentes rodoviários. 3.
Inteligência Artificial. 4. Detecção de padrões. 5.
Classificação de acidentes. 6. Predição de acidentes.
7. Redes Georreferenciadas. 8. HSM. I. Título.

FOLHA DE JULGAMENTO

Candidata: Engenheira **MARIA LIGIA CHUERUBIM.**

Título da tese: "Possibilidades em inteligência artificial na detecção de padrões e previsão de acidentes em rodovias".

Data da defesa: 08/03/2019.

Comissão Julgadora:

Prof. Associado **Irineu da Silva**
(Orientador)
(Escola de Engenharia de São Carlos/EESC)

Prof. Associado **Paulo César Lima Segantine**
(Escola de Engenharia de São Carlos/EESC)

Profa. Dra. **Barbara Stolte Bezerra**
(Universidade Estadual Paulista "Júlio de Mesquita Filho"/UNESP – Bauru)

Prof. Dr. **Alneu de Andrade Lopes**
(Instituto de Ciências Matemáticas e de Computação/ICMC-USP)

Prof. Dr. **Edson Augusto Melanda**
(Universidade Federal de São Carlos/UFSCar)

Coordenadora do Programa de Pós-Graduação em Engenharia de Transportes:

Profa. Associada **Ana Paula Camargo Larocca**

Presidente da Comissão de Pós-Graduação:
Prof. Titular **Murilo Araujo Romero**

Resultado:

Aprovado

Aprovada

APROVADA

Aprovado

Aprovado

DEDICATÓRIA

Dedico esse doutorado a todos os homens e mulheres que lutam pela construção de um mundo com mais amor, respeito e igualdade.

Aos meus amados pais, Izildinha (Dona Nenê) e Getúlio, e a minha irmã Fernanda, que sempre me incentivaram a realizar meus sonhos, tornando-os seus também.

EPÍGRAFE

*“When we see the Earth from space we see ourselves as a whole. We see the unity, but not the divisions. One planet, one human race. We're not the same, but we're one We are here together and we need to live together with tolerance and respect. Our only boundaries are the way we see ourselves, the only borders the way we see each other. We must become global citizens. Our voices are important. We give our elected officials their power, but we can take it away. We at all time travelers, traveling together into the future. But let us make that future a place we want to visit.
Be brave
Be determined
Overcome the odds
It can be done.”*

Stephen Hawking - “Global Citizen”.

RESUMO

CHUERUBIM, M. L. *Possibilidades em Inteligência Artificial na detecção de padrões e previsão de acidentes em rodovias*. 246p. Tese (Doutorado) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2019.

Nesta tese é apresentada uma investigação de possibilidades em Inteligência Artificial na detecção de padrões e previsão de acidentes em rodovias. Para tanto, é realizada uma avaliação de diferentes técnicas de Aprendizado de Máquina baseadas em abordagens de agrupamento, classificação e predição de links, com base em dados de acidentes georreferenciados e modelados em estruturas de clustering, árvores (CART) e redes (redes neurais artificiais, redes bayesianas e redes complexas). Os resultados revelaram que as abordagens baseadas em redes complexas possibilitaram a detecção de estruturas de agrupamentos mais robustas, quando comparadas as técnicas tradicionais de clustering, uma vez que consideram na estrutura topológica dos dados conceitos de vizinhança. O agrupamento dos dados proporciona a redução da heterogeneidade das bases de dados, bem como a obtenção de regras de decisão (CART) com maior probabilidade de ocorrência e taxa geral de acerto. A classificação supervisionada da severidade dos acidentes em redes, utilizando as modelagens de redes neurais artificiais e redes bayesianas, permitiu identificar simultaneamente os fatores contribuintes a ocorrência dos acidentes, sejam estes associados ao motorista, as variáveis de infraestrutura viária ou as condições do ambiente. No entanto, por considerar o peso das variáveis utilizadas no processo de modelagem, a classificação por redes bayesianas tende a ser mais realista, sendo menos sensível ao *overfitting*. Quanto à predição dos acidentes, foi possível pela predição de links utilizando a abordagem de redes complexas bipartidas, identificar alta correlação entre os acidentes preditos e os acidentes observados para uma determinada época. A abordagem proposta é flexível ao número de variáveis necessárias ao processo de modelagem, o que permite a realização de um diversificado número de estudos. No entanto, quando se considera a modelagem simplificada, formada pelas variáveis recomendadas pelo *Highway Safety Manual* (HSM), da *Association of State Highway na Transportation* (AASHTO), verifica-se que a predição é mais precisa e acurada, uma vez que esta modelagem considera fundamentalmente variáveis de infraestrutura viária, enquanto que na modelagem geral são consideradas também variáveis ambientais, que são mais variantes no tempo e no espaço. Até o presente momento, o método proposto é o mais adequado para explicar o comportamento e aspectos dinâmicos em ambientes rodoviários. No entanto, a abordagem proposta foi limitada pela quantidade de dados explorados, bem como por anomalias decorrentes a processos de execução de obras no trecho da rodovia em análise. Ademais, pode ser aplicada em problemas de diferentes escalas e para diversos estudos de caso. Portanto, por meio da modelagem de redes bipartidas georreferenciadas é possível não apenas realizar a predição de acidentes em rodovias, como também verificar a variação da acidentalidade viária e os níveis de segurança e desempenho de uma rodovia.

Palavras-chave: banco de dados georreferenciados; acidentes rodoviários; métodos de agrupamento; métodos de classificação; predição de acidentes; Inteligência Artificial, Aprendizado de Máquina, HSM.

ABSTRACT

CHUERUBIM, M. L. *Possibilities in Artificial intelligence in the detection of patterns and prediction of accidents on highways*. 246p. Thesis (PhD) – School of Engineering of São Carlos, University of São Paulo, São Carlos, 2019.

This thesis presents an investigation of Possibilities in Artificial Intelligence in the detection of patterns and prediction of accidents on highways. For this, it is performed an evaluation of different machine learning techniques based on grouping, classification and prediction of links, based on data from georeferenced accidents and modeled on clustering structures, trees (CART) and networks (artificial neural network, Bayesian network and complex network). The results revealed that the approaches based on complex networks enabled the detection of structures of groupings more robust, when comparing traditional clustering techniques, when they consider in the structure topological data neighborhood concepts. The grouping of data provides the reduction of heterogeneity of the database, as well as obtaining decision rules (CART) with the higher probability of occurrence and general rate of hit. The classification supervised of the severity of accidents in networks, using the modelling of artificial neural networks and Bayesian networks, it allowed to simultaneously identify the factors contributing to the occurrence of accidents, whether these are associated with the driver, the road infrastructure variables or the environmental conditions. However, considering the weight of the variables used in the modeling process, the classification by Bayesian networks tends to be more realistic, being less sensitive too *overfitting*. Regarding the prediction of accidents, it was possible to predict links using the approach of complex bipartite networks, to identify high correlation between predicted accidents and accidents observed for a certain time. The proposed approach is flexible to the number of variables necessary for the modeling process, which allows the realization of a diversified number of studies. However, when you consider the “simplified modelling”, formed by the variables recommended by the Highway Safety Manual (HSM), of the Association of State Highway na Transportation (AASHTO), it is verified that the prediction is more accurate and precise, since this modelling considers fundamentally variables of road infrastructure, while that the “general modelling” are considered the also environmental variables, which are more variants in time and space. Up to the present moment, the proposed method is best suited to explain the behavior and dynamic aspects in the environment road. However, the proposed approach was limited by the amount of data explored, as well as by anomalies resulting from processes of execution of works in the stretch of the highway under analysis. Moreover, it can be applied to problems of different scales and for several case studies. Therefore, through the modeling of geo-referenced bipartite networks, it is possible not only to predict accidents on highways, but also to check the variation of road accidents and the safety and performance levels of a highway.

Keywords: georeferenced database; road accidents; grouping methods; classification methods; prediction of accidents; Artificial Intelligence, machine learning, HSM.

Lista de Figuras

Figura 2.1 - Redução de dados pela projeção de componentes principais	16
Figura 2.2 - Processo de clusterização de um conjunto de dados	20
Figura 2.3 - Processo de ajustamento do conjunto de treinamento de classes no processo de construção de árvores de decisão	24
Figura 2.4 - Representação do modelo de ANN com MLP	28
Figura 2.5 - Estrutura geral de uma rede bayesiana	29
Figura 2.6 - Balanceamento de classes de dados pela técnica undersampling	31
Figura 2.7 - Balanceamento de classes de dados pela técnica oversampling	31
Figura 2.8 - Processo de balanceamento de classes de dados pela técnica SMOTE	32
Figura 2.9 - Desempenho da classificação com ANN utilizando MPL em bases de dados de acidentes utilizando técnicas de amostragem	35
Figura 2.10 - Construção das redes complexas pelo método dos vizinhos mais próximos	40
Figura 2.11 - Detecção de comunidades em redes complexas	41
Figura 2.12 - Rede Bipartida	42
Figura 2.13 - Rede bipartida ponderada	44
Figura 2.14 - Evolução do número de publicações por ano obtidas a partir de cada uma das bases consultadas	52
Figura 2.15 - Restrição do número de trabalhos consultados com base na estratégia de busca	52
Figura 2.16 - Mapeamento da filiação dos autores cujos trabalhos foram seleccionados	53
Figura 2.17 - Nuvem de palavras obtidas a partir dos metadados dos 44 artigos seleccionados	54
Figura 2.18 - Área de concentração	58
Figura 2.19 - Problema de pesquisa	58
Figura 2.20 - Método de pesquisa	58
Figura 2.21 - Benefícios de desempenho	58
Figura 2.22 - Número de objetivos	59
Figura 2.23 - Posição da variável no modelo	59
Figura 2.24 - Número de variáveis utilizadas	59
Figura 3.1 - Localização do trecho em estudo - escala: 1:500.000	66
Figura 3.2 - Estação de referência no CCO da concessionária “Rota das Bandeiras”	67

Figura 3.3 - Antena do receptor GNSS remoto instalada na parte superior do veículo utilizado na coleta dos dados	67
Figura 3.4 - Padrões de barreiras de contenção	70
Figura 3.5 - Processo geral de integração de registros observacionais de acidentes rodoviários com informações geoespaciais	73
Figura 3.6 -Obtenção da “ <i>Base Prediction</i> ” por meio de técnicas de AM	74
Figura 3.7 - Proposta de modelagem de dados rodoviários com ênfase na predição de acidentes com base em redes complexas bipartidas	78
Figura 3.8 - Instanciação da rede e criação do conjunto de vértices V_1 e V_2	79
Figura 3.9 - Etapa de criação das arestas e topologia inicial da rede	80
Figura 3.10 - Predição de links em redes complexas bipartidas	81
Figura 3.11 - Ilustração de um conjunto de links promissores	82
Figura 4.1 - Variação do percentual da variância acumulada em função da redução do número de variáveis do banco de dados pela técnica de PCA	87
Figura 4.2 - Comunalidades das variáveis resultantes da PCA	88
Figura 4.3 - Gráfico com o número de componentes	88
Figura 4.4 - Clustering para a base com VM para 2009-2012	90
Figura 4.5 - Clustering para a base com VM para 2012-2016	91
Figura 4.6 - Clustering para a base com VNF para 2009-2012	92
Figura 4.7 - Clustering para a base com VNF para 2012-2016	93
Figura 4.8 - Clustering para a base com VF para 2009-2012	94
Figura 4.9 - Clustering para a base com VF para 2012-2016	95
Figura 4.10 - Construção das redes complexas para VM de 2009-2012 com a variação do parâmetro $k - NN$ e do número de comunidades (C)	97
Figura 4.11 - Comunidades em redes complexas considerando a base de VM para 2009-2012	98
Figura 4.12 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VM de 2009-2012	98
Figura 4.13 - Construção das redes complexas para VM de 2012-2016 com a variação do parâmetro $k - NN$ e do número de comunidades (C)	99
Figura 4.14 - Comunidades em redes complexas considerando a base de VM para 2012-2016	99
Figura 4.15 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VM de 2012-2016	100
Figura 4.16 - Construção das redes complexas para VNF de 2009-2012 com a variação do parâmetro $k - NN$ e do número de comunidades (C)	100

Figura 4.17 - Comunidades em redes complexas considerando a base de VNF para 2009-2012	101
Figura 4.18 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VNF de 2009-2012	101
Figura 4.19 - Construção das redes complexas para VNF de 2012-2016 com a variação do parâmetro $k - NN$ e do número de comunidades (C)	102
Figura 4.20 - Comunidades em redes complexas considerando a base de VNF para 2012-2016	102
Figura 4.21 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VNF de 2012-2016	103
Figura 4.22 - Construção das redes complexas para VF de 2009-2012 com a variação do parâmetro $k - NN$ e do número de comunidades (C)	103
Figura 4.23 - Comunidades em redes complexas considerando a base de VF para 2009-2012	104
Figura 4.24 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VF de 2009-2012	104
Figura 4.25 - Construção das redes complexas para VF de 2012-2016 com a variação do parâmetro $k - NN$ e do número de comunidades (C)	105
Figura 4.26 - Comunidades em redes complexas considerando a base de VF para 2012-2016	105
Figura 4.27 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VF de 2012-2016	106
Figura 4.28 - Variação da AUC para a C_1 considerando a base de severidade com VM para o período de “antes” (2009-2012)	117
Figura 4.29 - Análise das variáveis contribuintes a ocorrência dos acidentes com VM no período de “antes” (2009-2012) na comunidade C_2	118
Figura 4.30 - Variação da AUC para a C_3 considerando a base de severidade com VM para o período de “antes” (2009-2012)	119
Figura 4.31 - Análise das variáveis contribuintes a ocorrência dos acidentes com VM no período de “antes” (2009-2012) na comunidade C_1	120
Figura 4.32 - Variação da AUC para a C_2 considerando a base de severidade com VM para o período de “antes” (2009-2012)	121
Figura 4.33 - Análise das variáveis contribuintes a ocorrência dos acidentes com VM no período de “antes” (2009-2012) na comunidade C_3	122

Figura 4.34 - Variação da AUC para a C_4 considerando a base de severidade com VM para o período de “depois” (2012-2016)	123
Figura 4.35 - Análise das variáveis contribuintes a ocorrência dos acidentes com VM no período de “depois” (2012-2016) na comunidade C_5	124
Figura 4.36 - Variação da AUC para a C_1 considerando a base de severidade com VNF para o período de “antes” (2009-2012)	125
Figura 4.37 - Variação da AUC para a C_3 considerando a base de severidade com VNF para o período de “antes” (2009-2012)	127
Figura 4.38 - Variação da AUC para a C_1 considerando a base de severidade com VNF para o período de “depois” (2012-2016)	128
Figura 4.39 - Variação da AUC para a C_2 considerando a base de severidade com VNF para o período de “depois” (2012-2016)	129
Figura 4.40 - Variação da AUC para a C_3 considerando a base de severidade com VNF para o período de “depois” (2012-2016)	130
Figura 4.41 - Variação da AUC para a C_4 considerando a base de severidade com VNF para o período de “depois” (2012-2016)	131
Figura 4.42 - Predição do número de “acidentes totais” para o ano de 2010	137
Figura 4.43 - Predição do número de “acidentes fatais e feridos” para o ano de 2010	138
Figura 4.44 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2010 utilizando diferentes modelagens	139
Figura 4.45 - Predição do número de “acidentes totais” para o ano de 2011	140
Figura 4.46 - Predição do número de “acidentes fatais e feridos” para o ano de 2011	140
Figura 4.47 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2011 utilizando diferentes modelagens	141
Figura 4.48 - Predição do número de “acidentes totais” para o ano de 2012 no período de “antes”	142
Figura 4.49 - Predição do número de “acidentes fatais e feridos” para o ano de 2012 no período de “antes”	142
Figura 4.50 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2012 (antes) utilizando diferentes modelagens	143
Figura 4.51 - Predição do número de “acidentes totais” para o ano de 2012 no período de “depois”	144
Figura 4.52 - Predição do número de “acidentes fatais e feridos” para o ano de 2012 no período de “depois”	144

Figura 4.53 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2012 (depois) utilizando diferentes modelagens	145
Figura 4.54 - Predição do número de “acidentes totais” para o ano de 2013	146
Figura 4.55 - Predição do número de “acidentes fatais e feridos” para o ano de 2013	146
Figura 4.56 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2013 utilizando diferentes modelagens	146
Figura 4.57 - Predição do número de “acidentes totais” para o ano de 2014	147
Figura 4.58 - Predição do número de “acidentes fatais e feridos” para o ano de 2014	147
Figura 4.59 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2014 utilizando diferentes modelagens	148
Figura 4.60 - Predição do número de “acidentes totais” para o ano de 2015	149
Figura 4.61 - Predição do número de “acidentes fatais e feridos” para o ano de 2015	149
Figura 4.62 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2015 utilizando diferentes modelagens	150
Figura 4.63 - Predição do número de “acidentes totais” para o ano de 2016	151
Figura 4.64 - Predição do número de “acidentes fatais e feridos” para o ano de 2016	151
Figura 4.65 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2016 utilizando diferentes modelagens	152
Figura 4.66 – Variação percentual dos acidentes “totais” utilizando diferentes abordagens de predição para o período de “antes” e “depois” da concessão da SP-065	153
Figura 4.67 - Variação percentual dos acidentes “fatais e feridos” utilizando diferentes abordagens de predição para o período de “antes” e “depois” da concessão da SP-065	154
Figura 4.68 – Espacialização dos acidentes reais observados no ano de 2009 na rodovia SP-065 do km 125 ao km 145,500	156
Figura 4.69 – Espacialização dos acidentes “totais” para o ano de 2010 na rodovia SP-065 do km 125 ao km 145,500	157
Figura 4.70 – Espacialização dos acidentes “fatais e feridos” para o ano de 2010 na rodovia SP-065 do km 125 ao km 145,500	158
Figura 4.71 – Espacialização dos acidentes “totais” para o ano de 2011 na rodovia SP-065 do km 125 ao km 145,500	159
Figura 4.72 – Espacialização dos acidentes “fatais e feridos” para o ano de 2011 na rodovia SP-065 do km 125 ao km 145,500	160
Figura 4.73 – Espacialização dos acidentes “totais” para o ano de 2012 no período de “antes” na rodovia SP-065 do km 125 ao km 145,500	161
Figura 4.74 – Espacialização dos acidentes “fatais e feridos” para o ano de 2012 no período de “antes” na rodovia SP-065 do km 125 ao km 145,500	162

Figura 4.75 – Espacialização dos acidentes “totais” para o ano de 2012 no período de “depois” na rodovia SP-065 do km 125 ao km 145,500	163
Figura 4.76 – Espacialização dos acidentes “fatais e feridos” para o ano de 2012 no período de “depois” na rodovia SP-065 do km 125 ao km 145,500	164
Figura 4.77 – Espacialização dos acidentes “totais” para o ano de 2013 na rodovia SP-065 do km 125 ao km 145,500	165
Figura 4.78 – Espacialização dos acidentes “fatais e feridos” para o ano de 2013 na rodovia SP-065 do km 125 ao km 145,500	166
Figura 4.79 – Espacialização dos acidentes “totais” para o ano de 2014 na rodovia SP-065 do km 125 ao km 145,500	167
Figura 4.80 – Espacialização dos acidentes “fatais e feridos” para o ano de 2014 na rodovia SP-065 do km 125 ao km 145,500	168
Figura 4.81 – Espacialização dos acidentes “totais” para o ano de 2015 na rodovia SP-065 do km 125 ao km 145,500	169
Figura 4.82 – Espacialização dos acidentes “fatais e feridos” para o ano de 2015 na rodovia SP-065 do km 125 ao km 145,500	170
Figura 4.83 – Espacialização dos acidentes “totais” para o ano de 2016 na rodovia SP-065 do km 125 ao km 145,500	171
Figura 4.84 – Espacialização dos acidentes “fatais e feridos” para o ano de 2016 na rodovia SP-065 do km 125 ao km 145,500	172

Lista de Tabelas

Tabela 2.1: Desempenho da classificação por ANN com MPL e usando banco de dados desequilibrado	33
Tabela 2.2: Desempenho da classificação por ANN com MPL e usando banco de dados <i>undersampling</i>	33
Tabela 2.3: Desempenho da classificação por ANN com MPL e usando banco de dados <i>oversampling</i>	33
Tabela 2.4: Desempenho da classificação por ANN com MPL e usando banco de dados SMOTE	33
Tabela 2.5: Resultado da pesquisa a partir da Expressão de busca	51
Tabela 2.6: Processo de seleção dos trabalhos considerando o título, o resumo, a leitura do artigo e a verificação dos critérios de inclusão e exclusão	51
Tabela 2.7: Os dez principais artigos selecionados utilizando a métrica “Google Scholar”	54
Tabela 2.8: Taxonomia das categorias por dimensão para análise dos artigos	55
Tabela 3.1: Coordenadas geodésicas das estações da RBMC	68
Tabela 3.2: Coordenadas geodésicas das estações da RBMC	68
Tabela 3.3: Coordenadas UTM estimadas para a base de referência do CCO	69
Tabela 3.4: Representação genérica da base de dados	79
Tabela 4.1 - Variância total explicada	87
Tabela 4.2: Discriminação geral dos agrupamentos com VM em função dos fatores destoantes e dominantes	107
Tabela 4.3: Discriminação geral dos agrupamentos com VNF em função dos fatores destoantes e dominantes	109
Tabela 4.4: Discriminação geral dos agrupamentos com VF em função dos fatores destoantes e dominantes	110
Tabela 4.5: Redução da dimensionalidade do banco de dados por seleção, agregação e exclusão	111
Tabela 4.6: Variáveis discriminatórias e essenciais ao desenvolvimento da abordagem proposta	112
Tabela 4.7: Performance dos classificadores ANN e BN para bases de severidade de acidentes rodoviários	113
Tabela 4.8: Tempo de processamento do processo de classificação usando ANN e BN para bases de severidade de acidentes rodoviários	113
Tabela 4.9: Matriz confusão obtida com ANN para VM no período de 2009 a 2012	113
Tabela 4.10: Matriz confusão obtida com BN para VM no período de 2009 a 2012	113

Tabela 4.11: Matriz confusão obtida com ANN para VM no período de 2012 a 2016	114
Tabela 4.12: Matriz confusão obtida com BN para VM no período de 2012 a 2016	114
Tabela 4.13: Matriz confusão obtida com ANN para VNF no período de 2009 a 2012	114
Tabela 4.14: Matriz confusão obtida com BN para VNF no período de 2009 a 2012	114
Tabela 4.15: Matriz confusão obtida com ANN para VNF no período de 2012 a 2016	114
Tabela 4.16: Matriz confusão obtida com BN para VNF no período de 2012 a 2016	114
Tabela 4.17: Matriz confusão obtida com ANN para VF no período de 2009 a 2012	114
Tabela 4.18: Matriz confusão obtida com BN para VF no período de 2009 a 2012	114
Tabela 4.19: Matriz confusão obtida com ANN para VF no período de 2012 a 2016	114
Tabela 4.20: Matriz confusão obtida com ANN para VF no período de 2012 a 2016	114
Tabela 4.21: Exatidão do classificador ANN na modelagem da severidade da base de VM no período de 2009 a 2012	115
Tabela 4.22: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VM no período de 2009 a 2012	115
Tabela 4.23: Exatidão do classificador ANN na modelagem da severidade da base de VM no período de 2012 a 2016	115
Tabela 4.24: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VM no período de 2012 a 2016	115
Tabela 4.25: Exatidão do classificador ANN na modelagem da severidade da base de VNF no período de 2009 a 2012	115
Tabela 4.26: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VNF no período de 2009 a 2012	115
Tabela 4.27: Exatidão do classificador ANN na modelagem da severidade da base de VNF no período de 2012 a 2016	115
Tabela 4.28: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VNF no período de 2012 a 2016	115
Tabela 4.29: Exatidão do classificador ANN na modelagem da severidade da base de VF no período de 2009 a 2012	116
Tabela 4.30: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VF no período de 2009 a 2012	116
Tabela 4.31: Exatidão do classificador ANN na modelagem da severidade da base de VF no período de 2012 a 2016	116
Tabela 4.32: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VF no período de 2012 a 2016	116
Tabela 4.33: Frequência de acidentes “totais” para o período de “antes” e “depois” da concessão da rodovia SP-065	152

Tabela 4.34: Frequência de acidentes “fatais e feridos” para o período de “antes” e “depois” da concessão da rodovia SP-065 153

Lista de Acrônimos

AA	Adamic-Adar
AASHTO	American Association of State Highway and Transportation Officials
AM	Aprendizado de Máquina
ANN	Artificial Neural Network
AUC	Area Under the curve of the Receiver Operating Characteristic
<i>BC</i>	Betweenness Centrality
BIM	Building Information Model
BNN	Bayesian Neural Network
CART	Classification And Regression Tree
<i>CC</i>	Closeness Centrality
CCO	Centro de Controle Operacional
<i>CN</i>	Common Neighbors
<i>DC</i>	Degree Centrality
EB	Método Empírico de Bayes
FDK	Função de Densidade de Kernel
FP	False Positive
FN	False Negative
GE	Google Earth
GNSS	Global Navigation Satellite System
HSM	Highway Safety Manual
IA	Inteligência Artificial
QGis	Geographic Information System. Open Source
$k - NN$	k - Nearest Neighbors
LGO	Leica Geo Office
MPL	Multi-Layer Perceptron
<i>MSE</i>	Mean Squares Error
NI	No Injury
PCA	Principal Component Analysis
RA	Resource Allocation
RB	Rede Bipartida
RTM	Regression-To-the-Mean

SIG	Sistemas de Informações Geográficas
SIRGAS	Sistema de Referência Geocêntrico para as Américas
SGB	Sistema Geodésico Brasileiro
SLR	Systematic Literature Review
<i>SQRes</i>	Somatório dos Quadrados Residuais
SMOTE	Synthetic Minority Over-sampling TEchnique
SRTM	Shuttle Radar Topography Mission
RBMC	Rede Brasileira de Monitoramento Contínuo dos Sistemas GNSS
RINEX	Receiver Independent Exchange Format
ROC	Receiver Operating Characteristics
TN	True Negative
TP	True Positive
TRB	Transportation Research Board
UTM	Universal Transversa de Mercator
VDMA	Volume Diário Médio Anual de veículos
VDM	Volume Diário Médio Mensal de veículos
VF	Vítimas Fatais
VM	Vítimas Materiais
VNF	Vítimas Não Fatais
WI	With injury
<i>WCN</i>	Weighted Common Neighbors
WSS	Within Cluster Sum of Square

SUMÁRIO

1 INTRODUÇÃO	1
1.1 Motivação	3
1.2 Identificação do problema	6
1.3 Objetivos e metodologia	8
1.4 Contribuições do trabalho	12
1.5 Estrutura do trabalho	14
2 CONCEITOS FUNDAMENTAIS E TRABALHOS RELACIONADOS	15
2.1 Componentes Principais	15
2.1.1 Conceitos fundamentais	17
2.1.2 Contribuições e retenção de componentes	19
2.1.3 Interpretação e escores de cada componente principal	19
2.2 Clustering	20
2.2.1 Medida de desempenho da clusterização dos dados	21
2.3 Classificação de dados	23
2.3.1 Árvores de decisão	23
2.3.2 Redes Neurais Artificiais	25
2.3.3 Redes Neurais Bayesianas	29
2.3.4 Balanceamento de classes	30
2.4 Redes Complexas	35
2.4.1 Conceitos básicos	36
2.4.2 Medidas de centralidade	37
2.4.3 Medidas de desempenho de agrupamentos em redes	40
2.4.4 Redes bipartidas	42
2.4.5 Predição de links	44
2.4.6 Métricas de avaliação e validação	47
2.5 Trabalhos relacionados	48
2.6 Considerações Finais	61
3 METODOLOGIA PROPOSTA	63

3.1 Protocolo de coleta de dados em ambiente rodoviário	65
3.1.1 Área de estudo	65
3.1.2 Rastreamento do trecho selecionado	66
3.2 Proposta de modelagem de dados usando técnicas de AM com ênfase na segurança viária	73
3.2.1 Processo de geração da base de predição	73
3.2.2 Classificação da severidade dos acidentes rodoviários	75
3.2.3 Modelagem da base de dados georreferenciada em uma rede bipartida	76
3.2.4 Predição de links em redes complexas bipartidas	80
3.3 Considerações finais	83
4 EXPERIMENTOS E RESULTADOS	85
4.1 Análise da viabilidade da aplicação da técnica de PCA em dados multivariados aplicados à acidentalidade viária	86
4.1.1 Redução da dimensionalidade do banco de dados e identificação das variáveis discriminantes	86
4.2 Clustering aplicado à detecção de agrupamentos de vítimas no trânsito	88
4.2.1 Detecção de clusters de acidentes de trânsito com vítimas com danos materiais utilizando cenários antes e depois da intervenção	90
4.2.2 Detecção de clusters de acidentes de trânsito com vítimas não fatais utilizando cenários antes e depois da intervenção	92
4.2.3 Detecção de clusters de acidentes de trânsito com vítimas fatais utilizando cenários antes e depois da intervenção	94
4.3 Redes complexas homogêneas aplicadas à detecção de comunidades de vítimas de acidentes de trânsito	95
4.3.1 Detecção de comunidades de acidentes de trânsito com vítimas com danos materiais utilizando cenários antes e depois da intervenção	97
4.3.2 Detecção de comunidades de acidentes de trânsito com vítimas não fatais utilizando cenários antes e depois da intervenção	100
4.3.3 Detecção de comunidades de acidentes de trânsito com vítimas fatais utilizando cenários antes e depois da intervenção	103
4.4 Extração de regras de decisão a partir de agrupamentos de acidentes de vítimas no trânsito pela abordagem de CART	106
4.5 Exclusão de variáveis por agregação ou eliminação em função do grau de relevância e do contexto da temática abordada	110

4.6 Classificação de bases de severidade de acidentes rodoviários pela abordagem de redes neurais artificiais e redes bayesianas	112
4.6.1 Regras de classificação obtidas pela abordagem de redes bayesianas a partir das bases de severidade de acidentes rodoviários	117
4.7 Modelagem e predição de acidentes no formato tabela atributo-valor utilizando a abordagem de redes complexas bipartidas	136
4.8 Análise espacial das modelagens de predição de acidentes	155
4.8.1 Discussão dos resultados obtidos pelas abordagens de predição de acidentes	172
4.9 Considerações finais	173
5 CONCLUSÕES	177
5.1 Síntese do trabalho	177
5.2 Considerações finais e contribuições	180
5.3 Recomendações	181
Referências	183
Apêndice A	197
Apêndice B	200
Apêndice C	204
Apêndice D	211
Apêndice E	213
Apêndice F	215
Apêndice G	219
Apêndice H	224
Apêndice I	227
Apêndice J	241
Apêndice K	246

1 INTRODUÇÃO

Nos últimos anos, vários modelos de predição de acidentes rodoviários foram desenvolvidos. Todos com a finalidade de identificar os possíveis efeitos que os fatores relacionados ao comportamento do condutor, a geometria da rodovia, ao veículo e as condições do meio ambiente exercem sobre a ocorrência dos acidentes e na gravidade das lesões (Caliendo, Guida e Parisi, 2007).

Os modelos de predição permitem identificar, quantificar e mapear os pontos críticos de acidentes em rodovias, também denominados de “*hot spots*” (Anderson, 2009; Plug Xia e Caulfield, 2011; Thill, 2000). O mapeamento de áreas com alta densidade de acidentes é de fundamental importância para a elaboração de contramedidas, visando melhorar a segurança rodoviária, uma vez que os acidentes compreendem um fenômeno complexo, aleatório, não linear e multidimensional (Karlaftis e Vlahogianni, 2011; Teodorović, 2008).

Por representar um fenômeno complexo e de natureza multicausal, a acidentalidade viária não deve ser tratada trivialmente por modelos estatísticos tradicionais, como os derivados da distribuição binomial negativa, da distribuição de Poisson, os modelos de regressão generalizada, dentre outros (Karlaftis e Vlahogianni, 2011). Estas técnicas não permitem modelar a variação aleatória do número de acidentes, sendo inadequadas ao tratamento de um grande número de variáveis, insensíveis a detecção de outliers, ruídos e dados ausentes, características estas inerentes aos dados de acidentes (Caliendo, Guida e Parisi, 2007; Karlaftis e Vlahogianni, 2011).

O Manual de Segurança Viária (HSM - *Highway Safety Manual*), publicado em 2010 pela AASHTO (*American Association of State Highway and Transportation Officials*), recomenda a aplicação de uma série de funções de desempenho de segurança e o uso do método Empírico de Bayes (EB) com fatores de calibração, para prever o desempenho e os níveis de segurança de uma rodovia (AASHTO, 2010; Saha, Alluri e Gan, 2011). No entanto, este método implica em uma demanda significativa de dados detalhados de características da rodovia, para derivar os fatores de calibração (Saha, Alluri e Gan, 2015).

Esta particularidade torna a aplicação integral do HSM economicamente inviável, no que se refere a coleta e manutenção das variáveis de calibração, considerando questões como a extensão da malha rodoviária, a falta de controle e padronização no processo de geração de informações pelas concessionárias e órgãos competentes, a alteração do uso e

ocupação do solo e o processo de implantação de obras de melhoria de infraestrutura viária (Saha, Alluri e Gan, 2011; Saha, Alluri e Gan, 2015).

Assim sendo, a priorização das variáveis de calibração do HSM com base no impacto da previsão da frequência de acidentes é de fundamental importância. Este procedimento auxilia na identificação das variáveis mais influentes, como por exemplo, o volume de tráfego, para as quais os dados devem ser continuamente coletados e mantidos (Saha, Alluri e Gan, 2015). Contudo, a qualidade do processo de predição ainda fica restrita a precisão de exatidão do EB, que se baseia na regressão tradicional para média (*Regression-To-the-Mean* - RTM), (AASHTO, 2010; Saha, Alluri e Gan, 2011).

Até os dias atuais as agências americanas de transportes ainda enfrentam inúmeros desafios para implementação do HSM, uma vez que atender a todos os seus requisitos de dados é uma tarefa desafiadora em termos de tempo e custos (Saha, Alluri e Gan, 2015). No Brasil, esta condição se agrava em função da falta de investimentos em planejamento urbano e infraestrutura viária, o que culmina na crescente demanda por mobilidade, seja em relação ao deslocamento de pessoas ou de cargas, fatores que consequentemente contribuem ao aumento do número de acidentes.

Neste contexto, abordagens fundamentadas em técnicas de Aprendizado de Máquina (AM) derivadas de conceitos da Inteligência Artificial (IA) têm sido investigadas, com a finalidade de proporcionar não apenas a mineração de grandes e massivos volumes de dados, mas ainda o desenvolvimento de modelagens de previsibilidade, focadas não apenas no potencial de explicação dos modelos funcionais ou matemáticos, mas também na acurácia e precisão dos parâmetros modelados (Karlaftis e Vlahogianni, 2011).

A IA é um campo da ciência e da engenharia de computação que busca reproduzir o conhecimento atribuído a sistemas inteligentes, que geralmente estão associados a características humanas, como a solução de problemas, o raciocínio lógico, a visão, a compreensão e a aprendizagem (Facelli *et al.*, 2011). A AM compreende uma área da IA cujo objetivo é o desenvolvimento de técnicas computacionais, fundamentadas em tarefas de aprendizado (supervisionado, não supervisionado e semissupervisionado), e a construção de sistemas capazes de adquirir o conhecimento de forma automática (Mitchell, 1997; Monard e Baranauskas, 2005).

As tarefas de aprendizado são realizadas a partir de exemplos extraídos com base em um conjunto de dados. No aprendizado supervisionado, a partir de um vetor de características (valores dos atributos) e do rótulo das classes, são estimados novos valores para estes atributos, a partir de um algoritmo denominado de classificador ou indutor.

Quando os rótulos das classes são representados por valores discretos o problema é conhecido como classificação, e os dados podem ser modelados usando a arquitetura de dados em árvores ou redes (Facelli *et al.*, 2011). Caso os valores sejam contínuos, o problema se restringe a utilização de modelos de regressão (Rezende *et al.*, 2005).

No aprendizado não supervisionado, o classificador analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de acordo com algum critério de similaridade ou dissimilaridade, por meio de algoritmos de agrupamentos baseados em clustering ou redes. No clustering, os dados são segmentados em agrupamentos denominados clusters, enquanto que na abordagem de redes os agrupamentos recebem a denominação de comunidades. Em ambas as abordagens, procura-se agrupar observações semelhantes ou detectar padrões regulares nos dados e, posteriormente, determinar o que cada agrupamento significa no contexto do problema abordado (Monard e Baranauskas, 2005).

No aprendizado semissupervisionado emprega-se uma abordagem híbrida que possibilita o aprendizado a partir de dados rotulados e não-rotulados (Chapelle *et al.*, 2010). Esta abordagem é amplamente utilizada em problemas do mundo real, ou seja, que não se restringem ao rótulo das classes de dados, como em Engenharia de Transportes que, em geral, são utilizados bancos de dados que já existem ou foram elaborados para outros fins.

1.1 Motivação

Problemas complexos de Engenharia de Transportes como nas áreas de planejamento, projeto geométrico e controle de tráfego em rodovias, em geral, envolvem a utilização de parâmetros incertos, não lineares, de natureza dinâmica no tempo e no espaço. Essa problemática não trivial requer a investigação de abordagens sofisticadas, que mesclam conhecimentos, técnicas e metodologias derivados de diversas áreas do conhecimento.

Em comum, estas abordagens têm destacado a necessidade da utilização de técnicas baseadas em heterogeneidade espacial dos dados, que possibilitam a modelagem das possíveis relações entre as variáveis consideradas, com a finalidade de se detectar, com uma margem razoavelmente baixa de erro, efeitos regionais e locais em segmentos rodoviários, por intermédio de grandes e diversificadas bases de dados (*big data*), (LeSage e Pace 2009; Li *et al.*, 2013; Liu, Khattak e Wali, 2017; Yao, Loo e Yang, 2016; Zhang *et al.*, 2015).

A obtenção de dados significativos, também denominados de dados inteligentes, por meio de grandes conjuntos de dados, é um grande desafio para os pesquisadores de segurança viária. Em geral, estes dados estão desbalanceados em função da severidade dos acidentes, o que implica que apresentem poucas observações relacionadas aos acidentes com vítimas fatais (classe minoritária), quando comparadas as observações associadas a classe majoritária de dados, relacionada à ocorrência dos acidentes sem vítima (com danos materiais) ou com vítimas não fatais (leves, moderadas e graves).

Deste modo, as técnicas de AM representam um poderoso conjunto de ferramentas analíticas que vêm apresentando resultados promissores em diversas áreas de aplicação da Engenharia de Transportes, dentre as quais se destacam as indicadas a seguir (Karlaftis e Vlahogianni, 2011): na segmentação de dados por meio de técnicas de agrupamento, com a finalidade de reduzir a heterogeneidade do conjunto de dados de acidentes (Kumar e Toshniwal, 2016); na extração de regras de decisão e classificação da severidade das lesões de trânsito, com base na seleção e associação de variáveis oriundas de dados aleatórios e estratificados, usando a arquitetura de árvores binárias (Pande e Abdel-Aty, 2006) ou redes (Khoo e Ahmed, 2018; Fu *et al.*, 2018; Zhu *et al.*, 2017) e na estimativa do número de acidentes, considerando tradicionalmente o efeito das variáveis de tráfego e da geometria da estrada (Ferraz *et al.*, 2008).

Em Engenharia de Transportes a área de pesquisa mais recorrente se refere a classificação e predição de acidentes usando modelagens de dados em redes. Neste tipo de abordagem, o conjunto de dados é representado tipicamente no formato de uma rede, em que os vértices são designados pelas classes de dados e pelo vetor de atributos. Cada vértice ou nó desta rede corresponde ao conjunto de dados original, que podem ser rotulados ou não. Os vértices rotulados são utilizados no processo de aprendizagem, por autocorrelação entre os vértices vizinhos. Nesta arquitetura, as arestas ou links representam os relacionamentos entre os vértices.

Em geral, os processos de classificação de dados de acidentes usando a abordagem de redes, utilizam o conhecimento indutivo, com base em algoritmos de Redes Neurais Artificiais (ANN – *Artificial Neural Network*) ou modificados por Redes Bayesianas (BNN - *Bayesian Neural Network*), (Alejo *et al.*, 2013; Delen, Sharda e Bessonov, 2006; Wach, 2016; Zhu *et al.* 2017). Desta forma, dado um conjunto de exemplos de treinamento, um indutor gera como saída um classificador (também denominado de hipótese), (Braga, Carvalho e Ludermir, 2005), de forma que, dado um novo exemplo,

este classificador possa prever com maior precisão possível a sua classe (como, por exemplo, condição do tempo: bom, chuvoso, garoa, neblina e nublado).

No processo de predição usando redes (também denominado de análise de links), se utiliza o conhecimento transdutivo ou semissupervisionado, com base na utilização de dados rotulados e não rotulados. Nesta abordagem, os vértices rotulados são utilizados no processo de propagação de informações para os demais vértices existentes na rede. A análise de links permite a inferência de conhecimento útil por meio de vértices e links pré-existent na rede. Neste caso, as relações previstas ou links preditos representam a probabilidade futura de relacionamento entre um par de vértices distintos (Rezende *et al.*, 2005).

A predição de links é uma tarefa ainda emergente na área de Engenharia de Transportes, uma vez que a rede obtida pela modelagem, usualmente, é esparsa, o que implica que os vértices da rede apresentem um número de conexões baixo, considerando o número total de vértices da rede e o desequilíbrio dos conjuntos de dados. Nestas circunstâncias, torna-se uma tarefa de grande dificuldade estimar os links mais prováveis da rede analisada (Getoor e Diehl 2005). O contrário, caracteriza uma rede densa, em que o número de conexões em cada vértice é relativamente similar ao tamanho total da rede (Newman, 2010).

No entanto, a modelagem de dados em redes, usualmente, apresenta um maior poder de explicação do problema, podendo ser utilizada em diferentes aplicações em transportes. Esta representação permite a captura da estrutura topológica dos dados; a representação de dados por meio de estruturas hierárquicas pelo processo de divisão de uma rede em redes menores (subredes); a aplicação de diferentes medidas de centralidade, que permitem identificar os vértices mais influentes na rede e a forma como se dá o fluxo de informações entre os mesmos; o uso de inferência coletiva; a aplicação de análises baseadas em vizinhança (Macskassy e Provost, 2007), e a detecção de comunidades (Ting, Bin e Qi, 2016). Por fim, agregam o formalismo da teoria dos grafos, explorados mais recentemente no âmbito das redes complexas (Newman, 2003).

Considerando que as redes complexas representam uma abordagem relativamente nova e sofisticada de AM em Engenharia de Transportes, se identificam na literatura aplicações promissoras que destacam a utilização desta modelagem em estudos de detecção de comunidades em diferentes aplicações de transportes. Como forma de exemplificação, se destaca a elaboração de métodos de previsão de congestionamento

com base na agregação de veículos e a extração de regras de decisão por meio de registros agrupados de acidentes de trânsito (Lin *et al.*, 2014; Ting, Bin e Qi, 2016).

Contudo, o foco principal dos trabalhos analisados se restringiu a aplicação de redes complexas homogêneas, nas quais os vértices e as relações existentes são do mesmo tipo. Além disso, os trabalhos não contemplaram a análise precisa da modelagem de dados usando esta abordagem, direcionando seus esforços na investigação dos fatores que corroboram a ocorrência dos acidentes.

Não foram encontrados na literatura de Engenharia de Transportes trabalhos que explorassem redes complexas heterogêneas na predição de acidentes em rodovias, bem como a partição do conjunto de vértices em redes bipartidas.

1.2 Identificação do problema

A maioria das pesquisas em Engenharia de Transportes envolvendo análise de dados (ou ciência de dados – *data science*) se fundamenta na utilização de técnicas estatísticas tradicionais, como ferramentas de suporte à decisão e na modelagem de dados por meio de técnicas de AM (Karlaftis e Vlahogianni, 2011).

Embora as técnicas estatísticas tradicionais apresentem fundamentos matemáticos amplamente aceitos e disseminados pela comunidade científica, frequentemente falham quando se lidam com dados complexos e não-lineares (multidimensionais, em que cada variável é representada por um vetor de valores), como no estudo do fenômeno da acidentalidade viária (Karlaftis e Vlahogianni, 2011).

A eficácia da abordagem tradicional (ou poder explicativo) é restrita ao erro de predição da função de estimação, já que desconsidera, em geral, os testes de hipóteses subjacentes, a estabilidade de parâmetros, a causalidade e a distribuição de erros (Karlaftis e Vlahogianni, 2011). Logo, por meio destas técnicas, como a regressão, não é possível tratar de forma efetiva a não linearidade dos dados, bem como a detecção de possíveis outliers e ruídos (DeTienne *et al.*, 2003; Karlaftis e Vlahogianni, 2011).

Embora a precisão seja uma métrica importante em análise de dados, a mesma deve ser considerada em conjunto com diferentes métricas (como, por exemplo, a acurácia) e fatores (como o nível de complexidade, a adequação da modelagem ao problema, a interpretabilidade dos resultados, a transferibilidade, a adaptabilidade à mudanças espaço-temporais, a escalabilidade, dentre outros), de forma a permitir a avaliação apropriada do desempenho da modelagem (Karlaftis e Vlahogianni, 2011; Kirby *et al.*, 1997).

Na modelagem com técnicas de IA não é necessário estabelecer pré-requisitos entre as variáveis utilizadas, sendo a solução determinada por processos iterativos, enquanto que nos métodos estatísticos tem-se uma solução “fechada” (Warner e Misra, 1996). Neste caso é proposta uma modelagem de dados mais sofisticada, flexível ao número de entradas e/ou saídas de dados (sem para tanto aumentar o número de parâmetros), proporcionando o tratamento de grandes conjuntos de dados. Esta abordagem combina elementos de aprendizado, adaptação, evolução e lógica difusa para criar “modelos inteligentes” a partir dos dados existentes. Utilizam modelos não paramétricos e não-lineares, que podem tratar diretamente a não-linearidade dos dados, sendo sensíveis a detecção de outliers, dados ausentes ou ruidosos (Gupta e Lam, 1996; Principe *et al.*, 2000; Karlaftis e Vlahogianni, 2011).

Outra limitação dos modelos obtidos pelas técnicas estatísticas tradicionais refere-se a multicolinearidade, que corresponde ao alto grau de correlação entre duas ou mais variáveis independentes. Enquanto que nos modelos obtidos pelas técnicas de IA pode-se supor, no processo de modelagem que determinadas variáveis independentes não estejam correlacionadas (DeTienne *et al.*, 2003; Karlaftis e Vlahogianni, 2011).

Indiferente do tipo de aprendizado, a modelagem de acidentes deve considerar a representação proposicional (Tabela atributo-valor) entre o conjunto de dados, de forma que se possa realizar a manipulação adequada com base em algoritmos de AM. Os formatos relacional e atributo-valor apresentam as seguintes diferenças relativas à representação por redes: (i) uma base de dados apresenta uma representação relacional quando os dados são armazenados em diversas tabelas, que contém diferentes tipos de variáveis bem como as relações entre as mesmas, enquanto que os dados proposicionais são armazenados em uma única tabela; (ii) os dados proposicionais consistem em variáveis com estruturas idênticas, sobre as quais assume-se que sejam Independentes e Identicamente Distribuídas (IID), enquanto que os dados relacionais representam diferentes tipos de variáveis relacionadas entre si; e (iii) os dados relacionais podem representar autocorrelação, ou seja, possuem conexões e similaridades entre si (Khosravi e Bina; 2010; Berton, 2016). Entretanto, uma única tabela também pode conter informações relacionais, quando por exemplo, os dados não são normalizados, as informações são redundantes ou apresentam várias observações relacionadas entre si (Berton, 2016).

Neste trabalho os dados tabulares são representados por meio de redes, no formato tabela atributo-valor, em que não se pressupõe relações entre as instâncias ou linhas da tabela, pela abordagem de redes complexas.

Em termos práticos, as abordagens de modelagens de dados tradicionais de IA podem ser utilizadas em conjunto no tratamento de dados complexos e multidimensionais, como em aplicações em tempo real, que geram grandes e massivos volumes de dados. Neste contexto, as técnicas estatísticas tradicionais permitem quantificar as incertezas (confiabilidade) e os relacionamentos entre as variáveis, bem como extrair/investigar as causalidades em modelagens derivadas de IA. Enquanto que as técnicas de IA permitem o tratamento de grandes bases, não restringe o número de variáveis utilizadas na modelagem, tratam a não-linearidade dos dados e a multicolinearidade (Karlaftis e Vlahogianni, 2011).

1.3 Objetivos e metodologia

O objetivo geral desta tese foi analisar métodos de agrupamento, classificação e predição a partir de dados georreferenciados de acidentes rodoviários, com base em técnicas de AM identificadas na literatura e desenvolver, a partir da integração destas abordagens, desenvolver um protocolo de coleta de dados em rodovias e um novo método de investigação de acidentes, considerando propriedades ainda não exploradas em Engenharia de Transportes, como precisão e acurácia de predição, especialmente no que se refere a representação dos dados em redes.

Para a detecção de agrupamentos de acidentes foram aplicadas técnicas de aprendizagem não supervisionada, por meio de algoritmos de mineração de dados, com a finalidade de reduzir a dimensionalidade e a heterogeneidade do conjunto de dados, em redes complexas homogêneas.

A modelagem de predição foi proposta com base na utilização de algoritmos de aprendizado semissupervisionado, que exploram critérios de vizinhança e análise de links em redes complexas heterogêneas bipartidas, nas quais as conexões ocorrem entre vértices de tipos diferentes. Ressalta-se que nesta pesquisa a modelagem de previsão de dados de acidentes rodoviários considera apenas redes bipartidas e heterogêneas, com no máximo dois tipos de vértices.

A hipótese dessa investigação é que por meio da aplicação de diferentes técnicas de IA, seja possível estudar fenômenos complexos, aleatórios e dinâmicos em ambiente rodoviário, usando a modelagem de dados em redes, melhorando potencialmente

atividades não triviais em Engenharia de Transportes como na detecção e previsibilidade de acidentes em rodovias, a partir de dados inteligentes.

O objetivo geral pode ser subdividido nos seguintes objetivos específicos:

- estudar as diferentes modelagens de dados de acidentes rodoviários presentes na literatura;
- propor com base nos trabalhos relacionados um protocolo de coleta de dados em rodovias com foco em segurança viária;
- aplicar técnicas de mineração de dados na redução da dimensionalidade do conjunto de variáveis obtidas, por meio de funções de similaridade;
- detectar agrupamentos de acidentes a partir de funções de dissimilaridade e critérios de vizinhança em redes;
- investigar estratégias de regras de classificação dos acidentes em diferentes segmentos da rodovia, usando estruturas de dados em árvores e redes;
- melhorar o desempenho da classificação, a partir da utilização de técnicas de amostragem de dados;
- extrair regras de decisão com os potenciais fatores de risco com base nos algoritmos de classificação investigados;
- desenvolver uma modelagem de predição de acidentes com base na teoria de redes complexas;
- avaliar o método proposto em bases de dados georreferenciadas de acidentes rodoviários, por meio de observações registradas no período antes-depois da concessão da rodovia.
- comparar o desempenho do método proposto com o EB do HSM, em relação ao poder de generalização e ajuste dos dados ao modelo.

Para o desenvolvimento deste trabalho foi proposto um protocolo de coleta de dados em ambiente rodoviário, atendendo aos requisitos de análise de segurança viária identificados no estado da arte. O mesmo tem por finalidade agregar aos registros observacionais de acidentes, fornecidos pelas concessionárias ou órgãos fiscalizadores, os dados derivados de projetos de infraestrutura viária e os dados geoespaciais obtidos direta ou indiretamente pela utilização de diferentes geotecnologias (GNSS – *Global Navigation Satellite System*, imagens de satélite e ortofotos).

Considerando-se o conjunto de variáveis selecionadas pela técnica de PCA, o conjunto de dados pode ser dividido em três subconjuntos distintos, de acordo com a severidade dos acidentes, que são: vítimas materiais (VM), vítimas não fatais (VNF) e vítimas fatais (VF). Desta forma, os algoritmos de agrupamento baseados no aprendizado não supervisionado podem ser aplicados para reduzir a heterogeneidade dos dados e otimizar a extração de regras de decisão, utilizando algoritmos de classificação gerando modelos em árvores ou redes, via aprendizado supervisionado.

Neste trabalho, foram utilizadas técnicas distintas de agrupamento fundamentadas em clustering e redes complexas homogêneas, com o propósito de avaliar iterativamente o desempenho destas abordagens nos subconjuntos de dados de severidade considerando toda a estrutura inteligente dos sistemas de transportes rodoviários.

Em relação aos algoritmos de agrupamento, foram utilizados na detecção de clusters o algoritmo *k-means*, com base na medida de dissimilaridade da distância euclidiana e na detecção de comunidades, o algoritmo *FastGreedy* (Clauset *et al.*, 2004), aplicado a partir de redes complexas homogêneas, construídas com o algoritmo *k-NN* (*k-Nearest Neighbors*), considerando o critério de vizinhança dos *k*-vizinhos mais próximos, (Berton, 2016).

O desempenho dos algoritmos de agrupamento foi avaliado a cada iteração pelos coeficientes da silhueta (Rousseeuw, 1987) e da modularidade (Clauset *et al.*, 2004), respectivamente, na técnica de clustering e redes complexas. O desempenho também foi avaliado em relação ao somatório dos erros quadráticos (WSS - *Within-Cluster Sum of Square*) e do percentual de acerto ou nível de confiança na extração de regras de decisão (CART - *Classification And Regression Tree*), utilizando dados balanceados e estruturados em árvores binárias. Este procedimento aumenta a acurácia de classificação e reduz o número de regras.

No balanceamento dos subconjuntos de dados de severidade é utilizada a técnica de amostragem SMOTE (*Synthetic Minority Over-sampling TEchnique*), que gera um conjunto de dados sintéticos com base na distribuição dos dados reais (Li *et al.*, 2017; Sain e Wulan, 2015).

A análise das regras de decisão obtidas pela CART, considera as estruturas mais robustas de agrupamento detectadas pelas técnicas de clustering ou redes complexas usando os subconjuntos de dados de severidade. Com a obtenção das regras de decisão pode-se selecionar as variáveis de maior peso para a formação dos agrupamentos e

descartar as variáveis menos relevantes na formação das regras, obtendo-se um novo e reduzido banco de dados.

A partir das variáveis selecionadas na etapa de agrupamento e na extração de regras de decisão, pode-se realizar a classificação dos dados de acidentes em redes. Considera-se como variável dependente a severidade dos acidentes sem vítima (NI – No Injury) e com vítima (WI – With Injury), utilizando algoritmos de ANN e BNN, com base em dados balanceados pela técnica SMOTE.

O desempenho dos algoritmos de classificação por ANN e BNN é analisado em função da qualidade do ajuste dos dados ao modelo, expressa pela métrica AUC (*Area Under the curve of the Receiver Operating Characteristic*), (Mujalli *et al.*, 2016).

Na modelagem de predição de acidentes, considera-se as variáveis selecionadas na etapa de agrupamento e extração de regras de decisão.

Com a finalidade de comparar os resultados obtidos com o EB do HSM, os dados podem ser modelados de duas formas distintas. Na primeira, se considera o número de acidentes totais (WI e NI), enquanto que na segunda somente os acidentes WI.

Independente da modelagem adotada, a predição é realizada pela análise de links utilizando abordagem de redes complexas heterogêneas bipartidas. Na instanciação do modelo são criados dois conjuntos de vértices distintos. O primeiro conjunto de vértices é definido pelos acidentes e seus respectivos atributos, enquanto que o segundo corresponde a localização destes acidentes. As arestas ou links são definidas pela relação entre os acidentes e a sua respectiva localização, ou seja, conectam vértices de tipos distintos (Newman, 2010).

Os relacionamentos descritos pelas arestas podem ser ponderados, utilizando alguma medida de similaridade. Neste trabalho, o peso das arestas foi definido com base na distância euclidiana calculada entre os vértices analisados.

Na predição de links em redes é utilizado o algoritmo *2-Hop Common Neighbor* (Lü e Zhou, 2010). Os links são preditos com uma pontuação (score) de acordo com a medida de similaridade adotada. Todos os possíveis links são preditos e ranqueados em ordem decrescente de probabilidade de ocorrência. Os links com maior pontuação com base na abordagem explorada, correspondem aos acidentes com maior probabilidade de ocorrência em determinados locais da rodovia.

A convergência do algoritmo ocorre quando se verifica a estabilização do score de predição. Gráficamente é indicado por um ponto de forte inclinação do gráfico, em que os links possíveis se localizam a esquerda deste ponto e os ruídos a direita. A qualidade

da solução é expressa pela precisão de predição, calculada pela razão entre os links preditos corretamente pela técnica (verdadeiros positivos) em relação ao total de links preditos (somatório dos verdadeiros positivos e falsos positivos).

Os locais com a maior frequência de acidentes preditos são denominados de pontos críticos. O mapeamento dos pontos críticos de acidentes pode ser realizado por meio de ferramentas de Sistemas de Informações Geográficas (SIG), com a utilização da Função de Densidade de Kernel (FDK).

Os resultados obtidos com a abordagem de predição de acidentes são validados considerando o EB do HSM. Para tanto, são calculados os resíduos acumulados em função do fluxo de tráfego de veículos observados em cada localização em que foram preditos acidentes. A magnitude dos resíduos indica o quão próxima ou divergente são estas diferentes soluções.

As soluções podem ser analisadas graficamente por gráficos denominados de *cure plots*, (Hauer, 2014) que expressam a relação entre o acúmulo residual do EB em relação aos valores observados, ou seja, dados reais de acidentes registrados pelas concessionárias, com relação a duas vezes o desvio padrão para mais ou para menos.

1.4 Contribuições do trabalho

Este trabalho deu origem a um protocolo de coleta de dados em rodovias, considerando requisitos de análise de dados em segurança viária, e ao desenvolvimento de uma modelagem de predição de acidentes, fundamentada na teoria de redes complexas. Ademais, permitiu as seguintes publicações e recomendações.

- Chuerubim, M. L. Previsibilidade de Acidentes de Trânsito com técnicas de Machine Learning. In: Data Science and Artificial Intelligence: Universidade de São Paulo, FEA/USP, novembro, 2018.
- Chuerubim, M. L.; Valejo, A.; Silva, I. Spatial and temporal identification of community structures of road accidents through homogeneous complex networks and measures of centrality. In: *8º Congresso Luso-Brasileiro para o Planeamento Urbano, Regional, Integrado E Sustentável, Cidades e Territórios - Desenvolvimento, atratividade e novos desafios*, PLURIS 2018, pp. 1-12, Coimbra, Portugal, 2018.
- Chuerubim, M. L.; Valejo, A. ;Silva, I. Bayesian Networks Modeling in Traffic Accidents: case study for Brazil. In: *32º Congresso de Pesquisa e Ensino em Transporte da ANPET*, 2018, Gramado. 32º Congresso de Pesquisa e Ensino em Transporte da ANPET, pp. 1-12, Gramado, Brasil, 2018.

- Chuerubim, M. L.; SILVA, I. Relação das ocorrências de acidentes de trânsito com as características do traçado da rodovia com base em análise de cluster. In: *20º Reunião de Pavimentação Urbana*, RPU, 2017, pp. 1-15, Florianópolis, Brasil, 2017.
- Lazaro, B. O.; Chuerubim, M. L. Análise descritiva de acidentes de tráfego em rodovias federais e sua relação com os padrões de uso e ocupação do solo circunjacente. In: *IV Congresso Latino-americano de Segurança no Trânsito e Formação de Condutores*, ALACI, pp. 43-44, Belo Horizonte, Brasil, 2018.
- Chuerubim, M. L.; Silva, I. Segmentação dos acidentes de trânsito em rodovias por meio de classes latentes utilizando a técnica de Análise de Cluster. In: *IV Congresso Latino-americano de Segurança no Trânsito e Formação de Condutores*, ALACI, Belo Horizonte, pp. 97-98, Brasil, 2018.
- Diban, G. M. F.; Bezerra, B. S.; Chuerubim, M. L.; Silva, I. Proposição de um banco de dados de acidentes para rodovias. In: *32º Congresso de Pesquisa e Ensino em Transporte da ANPET*, ANPET, pp. 1-2, Gramado, Brasil, 2018.
- Chuerubim, M. L.; Silva, I. Analysis of the viability of applying the Principal Components technique in multivariate data from traffic accidents. In: *Sigma Journal of Engineering and Natural Sciences*, 2018.

Além das publicações citadas, outras contribuições foram alcançadas. Na seção 2.5 é apresentada uma revisão sistemática da literatura. Acredita-se que este estudo pode auxiliar no desenvolvimento de pesquisas futuras, pois proporciona uma visão geral da área, na qual se considera as principais modelagens de dados utilizadas em segurança viária.

Na seção 3.1, é apresentado um protocolo de coleta de dados em ambiente rodoviário, considerando as variáveis mais recorrentes na literatura e dados adquiridos por diferentes geotecnologias. Espera-se que o protocolo proposto auxilie na elaboração de pesquisas em que o acesso a dados necessários a modelagem é restrito.

Nas seções 4.2 e 4.3, avalia-se o desempenho de métodos de detecção de agrupamentos em bases de acidentes rodoviários, por meio de métodos de clustering e comunidades em redes. O método de detecção de comunidades baseado em análise de vizinhança, apresenta o melhor desempenho no que se refere a generalização dos dados, pela assimilação de um maior número de variáveis e resulta em regras de decisão com maior acurácia e probabilidade de ocorrência, usando a estrutura de dados em árvores (CART), conforme descrito na seção 4.4.

Na seção 4.6, é realizada a comparação entre métodos de classificação em redes a partir de dados amostrados. É destacada a melhora da precisão geral de classificação, utilizando a técnica de sobreamostragem de dados SMOTE, bem como a performance dos algoritmos de classificação de ANN e BNN. Os resultados indicam que a BNN é mais precisa que a ANN, pois considera a função de probabilidade dos dados, sendo ideal a realização de estudos causais.

Por fim, na seção 4.7, é proposta uma nova abordagem de predição de acidentes rodoviários, que visa obter resultados tão confiáveis quanto o método EB do HSM, pela predição de links em redes complexas heterogêneas bipartidas. Neste capítulo, são discutidas diferentes modelagens de dados, bem como a viabilidade da aplicação do método proposto em engenharia de transportes.

1.5 Estrutura do documento

O restante deste documento está organizado da seguinte maneira:

Capítulo 2: neste capítulo são apresentados os conceitos fundamentais relacionados as técnicas de AM exploradas no âmbito da segurança viária, bem como suas formulações matemáticas, com ênfase em tarefas de redução da dimensionalidade da base de dados (PCA), agrupamento (clustering e redes complexas), classificação (CART, ANN e BNN) e predição de acidentes por meio da análise de links em redes complexas bipartidas.

Capítulo 3: neste capítulo é apresentado um protocolo de coleta de dados em rodovias, direcionado a aplicações de Engenharia de Transportes, bem como um método de modelagem de bases de dados de acidentes rodoviários, cuja contribuição central é a predição de acidentes usando a arquitetura de redes e princípios de análise de vizinhança, e correlações espaciais e temporais.

Capítulo 4: neste capítulo são descritos os experimentos e os resultados que corroboram aos problemas de pesquisa deste trabalho.

Capítulo 5: neste capítulo são apresentadas as considerações finais, bem como as conclusões e principais limitações detectadas no desenvolvimento da pesquisa, além de recomendações à trabalhos futuros.

2 CONCEITOS FUNDAMENTAIS E TRABALHOS RELACIONADOS

Neste capítulo são introduzidos conceitos e técnicas de AM direcionadas à modelagem de dados de acidentalidade viária. Na seção 2.1 é apresentada a técnica de redução da dimensionalidade da base de dados por PCA. Na seção 2.2 são discutidas as performances dos métodos de agrupamento de dados baseados no aprendizado não supervisionado e semissupervisionado, com ênfase em medidas de dissimilaridade, por clustering, e pela análise de vizinhança em redes complexas.

Na seção 2.3 são apresentadas algumas definições sobre classificação de dados pela aprendizagem supervisionada, usando estruturas de dados em árvores e redes (ANN e RB). É realizada uma breve explanação sobre métodos de balanceamento ou amostragem de dados (subamostragem, sobreamostragem e mistas).

E, por fim, na seção 2.4 são explicados os conceitos que estruturam a teoria de redes complexas, detecção de comunidades e predição de links, com ênfase em segurança viária.

2.1 Componentes Principais

Dentre as técnicas de mineração de dados utilizadas para a redução da dimensionalidade da base de dados, se destaca a PCA. É uma técnica estatística multivariada que permite analisar as relações existentes entre uma série de variáveis estatísticas, minimizando possíveis redundâncias e possibilitando a captação da maior variabilidade dos dados em poucos componentes, simplificando o mapeamento das informações.

Trata-se de um método de redução de dados, em que a máxima variância do conjunto de variáveis pode ser explicada pela classificação dos autovetores associados aos maiores autovalores da matriz de correlação, o que possibilita que a amostra de dados originais possa ser analisada a partir de um número pequeno de componentes independentes entre si e ortogonais.

A aplicação da técnica de PCA se inicia com o cálculo dos autovalores e de seus autovetores correspondentes, em função da matriz de covariâncias ou da matriz de correlações entre as variáveis. Os autovetores correspondem às componentes principais que representam as variáveis originais em cada um dos eixos e expressam a medida relativa da importância de cada variável em relação as componentes principais. Os sinais

podem ser positivos ou negativos e indicam, respectivamente, se a relação entre as variáveis é direta ou inversamente proporcional.

Os autovalores correspondem a soma, em coluna, de cargas fatoriais ao quadrado de um fator também conhecido como matriz latente e representa a quantia de variância explicada por esse fator. As cargas fatoriais representam a correlação entre as variáveis originais e os fatores, possibilitando o entendimento da natureza de cada fator.

A matriz de carregamentos de cada variável nas componentes principais, ao ser multiplicada pela matriz original de dados, resultará na matriz de contagens (escores) de cada caso em relação às componentes principais. Os escores podem ser dispostos em diagramas de dispersão, nos quais os eixos representam as duas componentes mais importantes indicando o grau de relacionamento entre os casos condicionados pelas variáveis medidas (Abdi e Williams, 2010).

O conjunto de n variáveis geram n componentes principais, que são obtidas em ordem decrescente de máxima variância (Johnson e Wichern, 1999; Hongyu, Sandanielo e Oliveira Junior, 2015). Portanto, embora a informação estatística presente nas n variáveis originais seja a mesma das n componentes principais, em geral, é comum obter em apenas 2 ou 3 das primeiras componentes principais mais de 90% desta informação, conforme ilustra a Figura 2.1.

Figura 2.1 – Redução de dados pela projeção de componentes principais.

Fonte: Autora (2018).

Esta técnica permite estabelecer índices de ordenação para um conjunto de n variáveis, a partir da definição de pesos e ponderações que traduzam a informação contida para um conjunto de m variáveis selecionadas. Além disso, a PCA pode ser utilizada na seleção das variáveis originais de maior peso na combinação linear das primeiras

componentes principais, que são as mais importantes do ponto de vista estatístico (Abdi e Williams, 2010).

2.1.1 Conceitos fundamentais

Algebricamente as componentes principais são combinações lineares de p variáveis originais $X_1, X_2, \dots, X_p$ e que apresentam médias e variâncias. As componentes principais são determinadas com base na matriz de covariâncias (S) ou na matriz de correlações R .

Em geral, parte-se da matriz R e utiliza-se a matriz S quando todas as variáveis originais têm escalas parecidas, já que a PCA não é invariante a escala. Na prática, extrair as componentes principais como autovetores de R é equivalente a calcular as componentes principais das variáveis originais, após cada uma ser padronizada de forma a terem variância unitária. A padronização deve ser feita quando as unidades de medidas das variáveis observadas não forem as mesmas (Mardia, Kent e Bibby, 1979).

Ressalta-se, que raramente há alguma correspondência entre as componentes obtidas a partir de R e S , havendo diferenças consideráveis entre seus desvios padrões. Portanto, o usuário deve conhecer a característica das variáveis do seu banco de dados para definir qual das duas matrizes irá adotar em suas análises, uma vez que a escolha da utilização das mesmas não se trata de um processo arbitrário.

A equação característica da matriz R ou S é a seguinte (Johnson e Wichern, 2002):

$$\det[R - \lambda I] = 0 \text{ ou } |R - \lambda I| = 0 \text{ e,} \quad (1)$$

$$R = \begin{bmatrix} 1 & r(X_1, X_2) & \cdots & r(X_1, X_p) \\ r(X_2, X_1) & 1 & \cdots & r(X_2, X_p) \\ \vdots & \vdots & 1 & \vdots \\ r(X_p, X_1) & r(X_p, X_2) & \cdots & 1 \end{bmatrix}. \quad (2)$$

A matriz R tem posto completo se o seu posto for igual p , ou seja, caso apresente o número máximo de linhas ou colunas linearmente independentes. A equação (1) terá p raízes chamadas de autovalores ou raízes características da matriz R . Desta forma, sendo $\lambda_1, \lambda_2, \dots, \lambda_p$ as raízes da equação (1), tem-se:

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p. \quad (3)$$

Para cada autovalor λ_i haverá um autovetor $\tilde{a}_i$ do tipo:

$$\sum_{j=1}^p a_{ij}^2 = 1 \text{ com } \tilde{a}_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ip} \end{bmatrix}. \quad (4)$$

Os autovetores $\tilde{a}_i$ são normalizados, ou seja, a soma dos quadrados dos coeficientes é igual a 1 e são ortogonais entre si. Por esta razão, apresentam as seguintes propriedades:

$$\sum_{j=1}^p a_{ij} \cdot a_{kj} = 1 \text{ com } (\tilde{a}_i' \cdot \tilde{a}_k) = 1, \quad (5)$$

$$\text{Sendo, } \sum_{j=1}^p a_{ij} = 0 \text{ com } (\tilde{a}_i \cdot \tilde{a}_k) = 0 \text{ para } i \neq k. \quad (6)$$

Desta forma, sendo $\tilde{a}_i$ o autovetor correspondente ao autovalor λ_i , então a i -ésima componente principal será obtida pela seguinte combinação linear:

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{ip}X_p, \quad (7)$$

ou, $Y = AX$, cuja variância $\hat{\sigma}_Y^2$ deve ser maximizada e está sujeita a $\sum_{i,j=1}^p a_{ij}^2 = 1$.

Neste caso, $Y_1, Y_2, \dots, Y_p$ são componentes principais não correlacionadas e a_{ij} os coeficientes da combinação linear, que correspondem ao peso da j -ésima variável com o i -ésima componente principal, representados na matriz A de ordem p . Os Y_p observados são designados como *scores* das componentes principais.

As combinações lineares representam a seleção de um novo sistema de coordenadas, obtido pela rotação do sistema original considerando original com eixos $X_1, X_2, \dots, X_p$. Os novos eixos $Y_1, Y_2, \dots, Y_p$ representam as direções com variabilidade máxima, permitindo uma interpretação mais simples da estrutura da matriz de covariância.

As componentes principais apresentam as seguintes propriedades (Johnson e Wichern, 2002):

- 1) A variância da componente principal Y_i é igual ao autovalor λ_i :

$$\hat{\sigma}(Y_i) = \lambda_i. \quad (8)$$

- 2) A primeira componente é a que apresenta a maior variância dentre as combinações possíveis, e assim sucessivamente:

$$\hat{\sigma}(Y_1) > \hat{\sigma}(Y_2) > \dots \hat{\sigma}(Y_p). \quad (9)$$

- 3) A variância total das variáveis originais é igual ao somatório dos autovalores que, por sua vez, é igual a variância total das componentes principais:

$$\sum \hat{\sigma}(X_i) > \sum \lambda_i > \dots \sum \hat{\sigma}(Y_i). \quad (10)$$

4) As componentes principais não são correlacionadas entre si:

$$\hat{Corr}(Y_i, Y_j) = 0. \quad (11)$$

2.1.2 Contribuição e retenção de componentes

A contribuição (*Contr*) de cada componente principal Y_i é obtida pela razão entre a variância de Y_i e a variância total, expressa em porcentagem. Ela equivale a proporção da variância total explicada pela componente principal Y_i .

O valor de *Contr* possibilita avaliar quantas componentes serão necessárias para efetuar a análise assegurando a máxima variância entre os indivíduos. Para aplicações em diversas áreas do conhecimento, o número de componentes utilizadas tem sido aquele que acumula 70% ou mais de proporção da variância total (Solanas, Manolov e Leiva, 2011), conforme indicado a seguir:

$$Contr = \frac{\hat{\sigma}(Y_1) + \dots + \hat{\sigma}(Y_p)}{\sum_{i=1}^p \hat{\sigma}(Y_i)} \cdot 100 \geq 70\% \text{ onde } k < p. \quad (12)$$

2.1.3 Interpretação e escores de cada componente principal

Esta etapa é feita identificando-se o grau de influência que cada variável X_j tem sobre o componente Y_i , expressa pela correlação entre cada X_j e a componente Y_i , que está sendo interpretada. Assim sendo, a correlação entre X_j e Y_i pode ser expressa por:

$$\hat{Corr}(X_j, Y_i) = r_{X_j, Y_i} = a_{ij} \cdot \frac{\sqrt{\hat{\sigma}(Y_i)}}{\hat{\sigma}(X_j)} = \sqrt{\lambda_i} \cdot \frac{a_{ij}}{\sqrt{\hat{\sigma}(X_j)}} \quad (13)$$

Para analisar a influência de $X_1, X_2, \dots, X_p$ sobre Y_i deve-se verificar o peso de cada variável sobre a componente Y_i , dado pela seguinte expressão:

$$w_1 = \frac{a_{i1}}{\sqrt{\hat{\sigma}(X_1)}}, w_2 = \frac{a_{i2}}{\sqrt{\hat{\sigma}(X_2)}}, \dots, w_p = \frac{a_{ip}}{\sqrt{\hat{\sigma}(X_p)}}, \quad (14)$$

sendo w_i o peso de X_i (Johnson e Wichern, 1999).

Caso o objetivo da análise seja a obtenção de índices, o processo termina na Equação (14). Todavia, se o objetivo da análise é comparar ou agrupar indivíduos, a análise se estende para o cálculo dos escores de cada componente principal. Os escores são os

valores dos componentes principais. Após a redução de p para q dimensões, as q componentes principais serão os novos indivíduos e toda análise é feita utilizando-se os escores desses componentes (Johnson e Wichern, 1999).

2.2 Clustering

A análise de agrupamentos por clusterização compreende um conjunto de técnicas multivariadas, cuja principal finalidade é agregar objetos ou conjunto de variáveis em grupos denominados clusters, com base em suas dissimilaridades (Hair *et al.*, 2009).

Os métodos de clusterização de dados são divididos em hierárquicos e não hierárquicos. Nos hierárquicos, os elementos de um grupo tornam-se elementos do grupo superior a partir de uma série de decisões de agrupamento, que combinam as observações em uma estrutura hierárquica. Nos métodos não hierárquicos, em geral, define-se k grupos iniciais e, então, aloca-se as n observáveis aos k grupos. Nestes métodos, a alocação inicial é arbitrária e iterativamente busca-se a alocação ótima (Hair *et al.*, 2009). A Figura 2.2 ilustra por meio de diferentes cores a segmentação de dados em clusters.

Figura 2.2 – Processo de clusterização de um conjunto de dados.

Fonte: Autora (2018).

Neste trabalho será considerado o método não hierárquico de clusterização de dados usando o algoritmo k - *means*, que aloca cada um dos n elementos existentes em um dos k grupos previamente estabelecidos. Objetiva-se com este método minimizar a soma dos quadrados residuais dentro de cada grupo com a finalidade de aumentar a homogeneidade do mesmo, ao mesmo tempo em que se busca maximizar essa soma entre os grupos, aumentando a diferença entre eles (Johnson e Wichern, 2002). Ao final da distribuição dos n elementos nos k grupos, é calculada a Somatório dos Quadrados Residuais ($SQRes$) de cada grupo, conforme indicado na equação (15):

$$SQRes(g) = \sum_{i=1}^{n_g} (x_i - \bar{x}_g)^2, \quad (15)$$

onde, $g = 1, \dots, k$ e n_g corresponde ao número de elementos do g -ésimo grupo. Após o cálculo de todas as $SQRes$ dos k agrupamentos efetua-se o somatório de todos eles, conforme indicado na equação (16):

$$SQRes = \sum_{i=1}^k SQRes(i). \quad (16)$$

Neste caso, quanto menor o valor de $SQRes$ mais homogêneos serão os grupos. A cada iteração são recalculadas a média do grupo e a soma dos quadrados correspondentes. Caso a $SQRes$ diminua, isto significa que a movimentação de um objeto de um grupo para outro é mantida, caso contrário, o objeto retorna ao grupo original ou se move para outro grupo, iniciando-se uma nova iteração. Quando a $SQRes$ não mais diminui ou o máximo de iterações predefinidas for ultrapassado, o processo termina e os grupos formados são apresentados.

Dentre as desvantagens do método destaca-se a liberdade de escolha do número de clusters feita pelo usuário. Caso o algoritmo selecione um número muito pequeno de clusters, isso pode acarretar na junção de dois *clusters* naturais e, por sua vez, um número exacerbado de clusters pode acarretar que um cluster natural seja dividido erroneamente em dois grupos distintos.

Na prática, a análise de clustering é utilizada na redução da heterogeneidade do conjunto de dados, o que possibilita simplificar ou reduzir o número de variáveis necessárias para a modelagem de um determinado problema. Nesta pesquisa, a técnica de clustering foi utilizada de forma complementar a técnica de PCA, visando identificar relacionamentos até então subjetivos entre as variáveis pertencentes à base de dados.

2.2.1 Medida de desempenho da clusterização dos dados

Para avaliar o desempenho a posteriori de um algoritmo de clusterização como o k -means, podem ser utilizadas medidas internas que avaliam a qualidade das partições ou clusters, em função do nível de compactação ou separabilidade dos mesmos. Dentre estas medidas se destaca o coeficiente da Silhueta, que quantifica o grau de dissimilaridade entre as variáveis que compõem cada cluster, auxiliando na seleção do melhor número de agrupamentos (Newman, 2004a; Kaufman e Rousseeuw, 1991).

O coeficiente da Silhueta permite avaliar para cada variável (observação) a sua coesão em relação a cada cluster. Além disso, possibilita verificar se os clusters apresentam uma

estrutura natural de agrupamento, ou seja, se as variáveis se encontram bem agrupadas ou posicionadas dentro de cada cluster e o quão forte é esta estrutura de agrupamento.

Esta medida é calculada para cada variável considerada na formação dos agrupamentos com base na seguinte formulação (Clauset *et al.*, 2004):

$$S(i) = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad (17)$$

onde, S é o coeficiente da Silhueta calculado para uma determinada variável i ; a_i é a dissimilaridade média existente entre a variável i em relação as demais variáveis pertencentes ao mesmo cluster, e; b_i é a dissimilaridade média da variável i em relação as demais variáveis pertencentes ao cluster vizinho mais próximo.

A média do coeficiente da Silhueta para todas as variáveis i em relação aos dados analisados é obtida por:

$$S_{WC} = \frac{1}{N} \sum_{i=1}^N S(i), \quad (18)$$

onde, S_{WC} corresponde ao coeficiente médio obtido após o cálculo da Silhueta para cada variável analisada e N ao número total de variáveis analisadas.

Os coeficientes $S(i)$ e S_{WC} variam de -1 a 1, fornecendo uma interpretação da solução individual para a variável ou para o conjunto de variáveis, respectivamente. Valores negativos não são desejáveis pois indicam que a distância média das variáveis pertencentes ao mesmo cluster é maior que a distância média em relação aos outros clusters.

Quanto mais próximo a 1 for o valor do $S(i)$ para uma determinada variável, mais próximo a variável analisada está das demais variáveis do seu agrupamento e vice-versa. E, em relação ao S_{WC} , valores próximos a 1 indicam uma forte estrutura de agrupamento, ou seja, as variáveis pertencentes a um agrupamento estão bem próximas.

Na literatura, a interpretação deste coeficiente é sugerida por Kaufman e Rousseeuw (1991). De acordo com os autores, valores entre $0,71 \leq S(i) \leq 1,00$ sugerem que os clusters encontrados apresentam estrutura robusta de agrupamento; valores de $0,51 \leq S(i) \leq 0,70$ que os clusters apresentam uma razoável estrutura de grupos; valores entre $0,26 \leq S(i) \leq 0,50$ que os clusters obtidos apresentam fraca estrutura de agrupamentos e valores de $S(i) \leq 0,25$ que os clusters não apresentam estrutura de agrupamento. Em geral, quanto maior o número de clusters menor será o valor médio da

Silhueta, ou seja, mais distante um ponto pertencente a um cluster estará em relação aos clusters vizinhos.

2.3 Classificação de dados

A classificação compreende uma tarefa preditiva de aprendizado de máquina que possibilita o mapeamento de um conjunto de dados de entrada em função de um número finito de classes, necessárias a investigação e modelagem de um determinado problema. Desta forma, cada variável ou observação presente na base de dados é atribuída a um conjunto predefinido de classes. Os classificadores possibilitam investigar a existência ou não de correlação entre os atributos das classes de dados analisadas.

A modelagem dos dados é feita por intermédio dos conjuntos de teste e treinamento que possibilitam treinar o classificador utilizado, bem como estimar os novos valores para os atributos das classes das variáveis em análise, com base nos valores ou atributos de entrada. Na prática é encontrada uma função a partir dos dados de treinamento, também denominada de modelo ou hipótese, por meio da qual são preditos ou estimados os novos valores dos atributos das classes. Dentre os métodos de classificação existentes se destacam os baseados em estruturas de árvore e rede.

Como é um método baseado exclusivamente em classes, isto implica que seja extremamente sensível ao desbalanceamento das classes de dados (Mujalli, López e Garach, 2016). Esta é uma característica comum em aplicações nas quais os dados de um subconjunto de classes aparecem com uma frequência maior que os dados das demais classes. Os conceitos fundamentais sobre os métodos de balanceamento de classe são apresentados na seção 2.3.4.

2.3.1 Árvores de Decisão

Os classificadores baseados em árvores de decisão compreendem técnicas analíticas que permitem detectar automaticamente as melhores variáveis preditoras e seus respectivos limiares, por meio da extração de regras de decisão do tipo "*se* → *então*", que podem ser utilizadas para descobrir comportamentos que ocorrem dentro de um conjunto específico de dados (Abellán, López e De Oña, 2013; De Oña *et al.*, 2013, 2014; Kashani, Shariat-Mohaymany e Ranjbari, 2011).

A estrutura da árvore de decisão é construída de forma recursiva em que cada nó representa uma classe e os ramos os seus respectivos atributos, de acordo com um limiar ou regra de decisão. Cada regra tem início na raiz da árvore e se desenvolve pelos ramos

da árvore, com a finalização nos nós de decisão. Cada nó terminal ou folha especifica o valor esperado para cada variável.

As regras de decisão são disjuntas, ou seja, uma única regra é disparada quando um novo exemplo é classificado. Na prática, uma regra de decisão é uma implicação do tipo: se A então B , em que A representa um conjunto de condições. Cada condição é definida por uma relação do tipo $atributo_i = valor_i$, $atributo_i < valor_i$, $atributo_i > valor_i$, $atributo_i \leq valor_i$ e $atributo_i \geq valor_i$, na qual o $valor_i \in$ ao domínio do $atributo$ em análise.

Para tanto são utilizadas métricas para maximizar o escore de pureza de cada nó dentre as possíveis variáveis de entrada. Este procedimento é denominado de poda da árvore, possibilitando a redução do número de nós internos e, consequentemente, da complexidade da árvore, conforme ilustra a Figura 2.3.

Figura 2.3 – Processo ajustamento do conjunto de treinamento de classes no processo de construção de árvores de decisão.

Fonte: Autora (2018).

Desta forma, com base em uma estrutura binária, as variáveis são agrupadas em função da sua importância para descrever a variável alvo. Por meio da estrutura da árvore de decisão é possível detectar os nós que contêm o maior número possível de ocorrências relacionadas a cada categoria das variáveis independentes. Em função da validação cruzada entre as amostras de teste e treinamento verifica-se a estrutura que melhor se ajusta ao conjunto de dados, podando a árvore quando necessário e excluindo, neste processo, os nós ou ramos que pouco agregam ao processo de classificação da variável alvo, ou seja, que apresentam pouco valor preditivo.

Neste trabalho, optou-se em utilizar o algoritmo de construção de redes CART e a métrica Gini, que busca maximizar a homogeneidade dos nós em relação a variável

dependente (Pande e Abdel-Aty, 2006). Essa métrica atinge valores mínimos (zero) quando todos os casos em um nó se enquadram em uma única categoria.

Os critérios de redução de Gini medem a utilidade de cada variável no que se refere a sua contribuição, visando maximizar o processo de ramificação da árvore. O valor do processo de divisão pode ser medido da seguinte forma (Pande e Abdel-Aty, 2006; López *et al.*, 2016; Wei *et al.*, 2017):

$$valor = impureza(nó\ raiz) - \sum_{n=1}^N P(n) \times impureza(n), \quad (19)$$

onde a *impureza* do nó raiz denota a entropia ou medida de Gini para a impureza, ou seja, a não homogeneidade do nó raiz e $P(b)$ expressa a proporção de observações atribuídas a ramificação que contém o nó b (López *et al.*, 2016; Pande e Abdel-Aty, 2006).

A medida de impureza do nó pode ser definida da seguinte forma (López *et al.*, 2016; Pande e Abdel-Aty, 2006):

$$impureza(nó) = 1 - \sum_{i=1}^{classes} \frac{(número\ de\ classes\ no\ nó)}{(número\ de\ casos\ atribuído\ ao\ nó)} \quad (20)$$

Quanto menor o valor obtido pela equação (20) maior será o grau de pureza do nó, o que implica que seu valor seja maior para nós menos homogêneos. Desta forma, ao se desenvolver uma CART este critério deve ser aplicado recursivamente do nó raiz aos nós descendentes até que se atinja a homogeneidade ótima dos nós ou que um limiar pré-estabelecido seja atendido (López *et al.*, 2016; Pande e Abdel-Aty, 2006).

2.3.2 Redes Neurais Artificiais

A ANN (*Artificial Neural Network*) é uma técnica amplamente utilizada em problemas lineares e não lineares, pois não requer que pressuposições a-priori sejam estabelecidas entre as variáveis. É indicada no tratamento de fenômenos que não são bem conhecidos ou que são derivados de fatores multicausais, nos quais a abordagem analítica tradicional demandaria alto tempo computacional (Mussone, Ferrari e Oneta, 1999).

O princípio das ANNs é minimizar o erro médio quadrático (MSE - *Mean Squares Error*) fornecido pela equação (Chang, 2005):

$$MSE = \frac{1}{N \times K} \sum_{i=1}^N \sum_{j=1}^k (t_{i,j} - a)^2, \quad (21)$$

onde t e a são os parâmetros estimados e observados, respectivamente; K é o número de neurônios da camada de saída e N o número de dados do conjunto de teste (Chang, 2005).

Na modelagem de dados de acidentalidade viária a ANN, usualmente, é construída por múltiplas camadas (MLP - *Multi-Layer Perceptron*), pelo treinamento supervisionado com base no algoritmo *back-propagation*. A aplicação deste algoritmo se baseia na regra delta generalizada, descrita por duas fases: a fase para a frente (*forward*) e a fase para trás (*backward*), (Facelli *et al.* 2011).

Na fase *forward*, cada variável da base de dados é armazenada em um neurônio da rede. A ponderação inicial se dá pelas conexões ou relações existentes entre as variáveis da camada de entrada. Cada neurônio nessa camada aplica a função de ativação a sua entrada total e produz um valor de saída, que é utilizado como valor de entrada pelos neurônios da camada seguinte. Esse processo se dá iterativamente até que os neurônios da camada de saída produzam cada um seu valor de saída, que é então comparado ao valor desejado para a saída desse neurônio. A diferença entre os valores produzidos e desejados para cada neurônio da camada de saída expressa o erro do processo de modelagem dos dados (Facelli *et al.* 2011).

O valor do erro de cada neurônio da camada de saída é então utilizado na fase *backward*, no ajuste dos pesos de entrada. O ajuste ocorre da camada de saída até a camada intermediária. O ajuste dos pesos de uma MLP pelo algoritmo *back-propagation* pode ser obtido pela seguinte equação (Facelli *et al.* 2011):

$$w_{j,l}(t+1) = w_{j,l}(t) + \eta x^j \delta_l, \quad (22)$$

onde, $w_{j,l}$ representa o peso entre um neurônio l e o j -ésimo atributo de entrada ou saída do j -ésimo neurônio da camada anterior; δ_l o erro associado ao l -ésimo neurônio; x^j corresponde a entrada recebida por esse neurônio, ou seja, o j -ésimo atributo de entrada ou a saída j -ésimo neurônio da camada anterior; e η a taxa de aprendizado da modelagem.

Nesta modelagem apenas os valores dos erros dos neurônios de saída são conhecidos. Desta forma, deve-se estimar o erro para as camadas intermediárias. No algoritmo *back-propagation*, o erro de um neurônio de uma camada intermediária é estimado como a soma dos erros dos neurônios da camada seguinte, cujos terminais de entrada estão conectados a ele. A ponderação ocorre pelo peso atribuído a estas conexões.

Deste modo, o cálculo do erro dependerá da camada em que se encontra o neurônio, como mostra a equação (23) e a equação (24), (Facelli *et al.* 2011):

$$\delta_l = f'_a e_l, \text{ se } n_l \in \text{camada}_{saída}, \quad (23)$$

$$\delta_l = f'_a \sum w_{l,k} \delta_k, \text{ se } n_l \in \text{camada}_{intermediária}, \quad (24)$$

onde, n_l representa o l -ésimo neurônio; f'_a a derivada parcial da função de ativação do neurônio; e e_l o erro quadrático cometido pelo neurônio de saída quando sua resposta (y_q) é comparada a desejada ($\hat{f}_q$), definido por:

$$e_l = \frac{1}{2} \sum_{q=1}^k (y_q - \hat{f}_q)^2. \quad (25)$$

O ajuste dos pesos é definido pela f'_a , utilizando o gradiente descendente da função de ativação. Essa derivada mede a contribuição de cada peso no erro da rede para cada variável utilizada na análise. Quando o valor da f'_a é positivo para um dado peso, isto indica que o peso atribuído a uma determinada variável está provocando um aumento da diferença entre os valores obtidos e desejados na modelagem. Neste caso, a magnitude do peso deve ser reduzida para minimizar o erro obtido. E, caso contrário, quando o valor da f'_a for negativo, o peso deve ser aumentado, para que os valores obtidos se aproximem dos valores esperados (Facelli *et al.* 2011).

A Figura 2.4 apresenta a estrutura da ANN com MLP, geralmente utilizada na modelagem do banco de dados de acidentes, definida fundamentalmente por 3 camadas (entrada, oculta e saída):

Figura 2.4 - Representação do modelo de ANN com MLP.

Fonte: Autora (2018).

Na prática, a modelagem se inicia pela extração aleatória de dois subconjuntos de dados: o conjunto de treinamento com 70% dos dados e o conjunto de teste com 30% dos dados. Estes conjuntos são utilizados de forma iterativa na aprendizagem e na validação da ANN. Utiliza-se, em geral, a *cross-validation* sem repetição, de forma que ao final do processamento todo o conjunto de dados tenha sido verificado. Estes subconjuntos não devem ser muito grandes, caso contrário podem dificultar o processo de aprendizagem da ANN.

Além disso, os subconjuntos de treinamento e de teste devem ser o mais homogêneo possível, para cada classe de variáveis, pois o aprendizado da ANN se dá com base no maior número de combinações possíveis. Quando se utilizam muitas variáveis ou os dados não são distribuídos de forma homogênea, a qualidade do aprendizado da ANN também tende a decair. Portanto, na camada de entrada da ANN deve-se selecionar um menor número de variáveis, para que se possa ter melhores resultados tanto na extração dos dados quanto na distribuição das classes nos subconjuntos de teste e treinamento. Além disso, os valores das variáveis de entrada (preditores) devem ser normalizados.

Neste estudo, para a transferência de informações entre os neurônios da camada de entrada e a camada oculta, utilizou-se a função de ativação tangente hiperbólica, pois a seleção da arquitetura da rede foi realizada automaticamente. E para a transferência de informações entre a camada oculta e a camada de saída, foi utilizada uma função de ativação linear, uma vez que todas as variáveis independentes são categóricas.

2.3.3 Redes Neurais Bayesianas

A estrutura de uma Rede Neural Bayesiana (BNN - *Bayesian Neural Network*) é muito semelhante a estrutura de uma ANN. Porém, estas abordagens de modelagem de dados são diferentes no que se refere aos mecanismos de previsão e treinamento (Neal, 1996; Xie, Lord e Zhang, 2007).

Um modelo típico utilizado para representar a frequência de acidentes a partir da modelagem de dados usando redes neurais, fundamentalmente considera duas funções, conforme ilustra a Figura 2.5.

Fonte: Adaptado de Xie, Lord e Zhang (2007).

A função de transferência dos dados reais para a camada oculta (f_h) é representada, em geral, por uma função de ativação tangente hiperbólica ($\tanh$). Enquanto a função de transferência dos valores de f_h para a camada de saída (f_o) por uma função linear. Desta forma, supondo que um conjunto de dados formado por n acidentes de trânsito possa ser representado pela relação $Y = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, onde x_i corresponde aos valores reais de acidentes, e y_i a quantidade de acidentes que se deseja estimar no local i , o número previsto de acidentes para um dado local i pode ser obtidos pela seguinte equação (Xie, Lord e Zhang, 2007):

$$\hat{y}_i = f(x_i, \theta) = b_2 + \sum_{j=1}^M \left\{ w_2(j, k) \tanh \left(\sum_{k=1}^P w_1(i, k) x_{i,k} + b_1(j) \right) \right\}, \quad (26)$$

onde P corresponde ao número de neurônios da camada de entrada, M representa o número de neurônios da camada oculta, $b_1(j)$ e b_2 correspondem aos vieses das camadas de entrada e oculta, respectivamente; $w_2(j)$ corresponde aos pesos que conectam a camada oculta a camada de saída; $w_1(i,k)$ corresponde ao k -ésimo elemento da i -ésima entrada, sendo $x_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,n}\}$; θ ao vetor que contém todos os parâmetros da rede ($b_1(j)$, b_2 , $w_1(j,k)$, $w_2(j,k)$, $i = \{1, 2, \dots, n\}$, $j = \{1, 2, \dots, M\}$ e $k = \{1, 2, \dots, P\}$), (Xie, Lord e Zhang, 2007).

2.3.4 Balanceamento de classes

Um dos principais desafios enfrentados na classificação, consiste no desequilíbrio entre o número de observações e os atributos (categorias) das classes presentes na base de dados (Alejo *et al.*, 2013; Anand *et al.*, 1993; Piri, Delen e Liu, 2018). Em especial, no tratamento de problemas multiclases, como em Segurança Viária, em que essa particularidade pode afetar o desempenho dos algoritmos de classificação (Cateni, Colla e Vannucci, 2014; Krawczyk, 2016; Sain e Wulan, 2015; Zhou e Liu, 2006).

O desequilíbrio também denominado de desbalanceamento da base de dados ocorre quando existem classes majoritárias e minoritárias, ou seja, quando o número de instâncias de uma determinada classe é muito maior do que de outras classes (Alejo *et al.*, 2013; Loyola-González *et al.*, 2016; Roy *et al.*, 2018; Yang e Wu, 2006). Esse fenômeno ocorre frequentemente em problemas do mundo real como em banco de dados de acidentes rodoviários, nos quais a classe predominante corresponde as observações relacionadas aos acidentes sem vítimas e a classe menos frequente aos acidentes com vítimas fatais e não fatais.

Quando se lida com bancos de dados desequilibrados, o algoritmo tende a convergir mais lentamente e acaba por tornar o processo de aprendizagem tendencioso, negligenciando as classes menos recorrentes no banco de dados (Alejo *et al.*, 2013; Anand *et al.*, 1993; Roy *et al.*, 2018; Galar *et al.*, 2016; Li *et al.*, 2018; Sain e Wulan, 2015).

Na literatura, o problema de desequilíbrio de classes é minimizado pela aplicação de técnicas de balanceamento ou amostragem de banco de dados como a subamostragem (undersampling), a sobreamostragem (oversampling) e a SMOTE (*Synthetic Minority Over-sampling TEchnique*), (Li *et al.*, 2017; Maloof, 2003; Weiss, 2004; Sain e Wulan, 2015). Estas técnicas reduzem a quantidade de amostras da classe majoritária e aumentam

a magnitude das amostras da classe minoritária para realizar o processo de reequilíbrio da base de dados.

A *undersampling* reduz a população da classe majoritária e implica em perda de informação, com a eliminação de amostras pertencentes à classe dominante, podendo afetar a qualidade do processo de classificação, conforme ilustra a Figura 2.6 (Alejo *et al.*, 2013; Castellanos *et al.*, 2018). A *oversampling* equilibra a distribuição de classes pelo aumento da população da classe minoritária, por meio da replicação aleatória das amostras presentes nestas classes, conforme ilustra a Figura 2.7. Em geral, este procedimento não implica em qualquer perda de informação, uma vez que nenhuma amostra é descartada (Alejo *et al.*, 2013; Castellanos *et al.*, 2018; Cateni, Colla & Vannucci, 2014).

Figura 2.6 – Balanceamento de classes de dados pela técnica *undersampling*.

Figura 2.7 – Balanceamento de classes de dados pela técnica *oversampling*.

Fonte: Adaptado de Fawcett (2016).

Na técnica SMOTE, a classe minoritária é balanceada considerando cada uma de suas instâncias e a introdução de amostras sintéticas baseadas em critérios de vizinhança, como a distância euclidiana, usando o algoritmo de vizinho mais próximo (Bolón-Canedo *et al.*, 2014; Chawla *et al.*, 2002).

A Figura 2.8 ilustra basicamente a aplicação da técnica de balanceamento de classes usando o algoritmo SMOTE.

Figura 2.8 – Processo de balanceamento de classes de dados pela técnica SMOTE.

Fonte: Adaptado de Fawcett (2016).

As técnicas que utilizam *oversampling* e SMOTE, embora sejam as mais promissoras ao balanceamento de bases de dados, no processo de classificação, apresentam um maior custo computacional quando comparada as técnicas de *undersampling*, em especial quando são utilizados classificadores baseados em similaridades, como o vizinho mais próximo. Além disso, a amostragem superabundante pode acarretar o excesso de ajuste dos dados de treinamento ao modelo, processo também conhecido como *overtuning* ou *overfitting* (Salunkhe e Mali, 2016).

Como exemplo, a seguir, é demonstrado o impacto na classificação da severidade em base de acidentes rodoviários, pela abordagem da ANN com MPL, pela utilização de diferentes técnicas de amostragem de dados. Na Tabela 2.1, nota-se que ao se considerar um conjunto de variáveis originais, analisadas de forma exploratória no âmbito da segurança viária, ocorre *overfitting* do algoritmo ANN com MPL para a classe majoritária de dados, que corresponde aos acidentes sem vítimas (SV), com uma taxa de acerto de 97,9%, enquanto que para a classe minoritária de dados, acidentes com vítimas (CV) esta taxa é de apenas 14,2%, com uma precisão geral de classificação de 76,7%. A aplicação da técnica *undersampling* resulta em uma melhora na precisão geral dos dados (82,0%). Contudo, resulta em perda da informação da classe majoritária de dados, conforme se verifica na Tabela 2.2.

Tabela 2.1: Desempenho da classificação por ANN com MPL e usando banco de dados desequilibrado.

		Preditos		
Subconjuntos	Nível da lesão	SV	CV	(%) Acerto
Treinamento	SV	2.163	46	97,9
	CV	643	106	14,2
	(%) Total	94,9	5,1	76,7
Teste	SV	954	21	97,8
	CV	282	44	13,5
	(%) Total	95,0	5,0	76,7

Tabela 2.2: Desempenho da classificação por ANN com MPL e usando banco de dados *undersampling*.

		Preditos		
Subconjuntos	Nível da lesão	SV	CV	(%) Acerto
Treinamento	SV	28	7	80,0
	CV	10	26	72,2
	(%) Total	53,5	46,5	76,1
Teste	SV	15	0	100,0
	CV	5	8	61,5
	(%) Total	71,4	28,6	82,0

Tabela 2.3: Desempenho da classificação por ANN com MPL e usando banco de dados *oversampling*.

		Preditos		
Subconjuntos	Nível da lesão	SV	CV	(%) Acerto
Treinamento	SV	2.826	116	96,1
	CV	0	2.916	100,0
	(%) Total	48,2	51,8	98,0
Teste	SV	1.199	67	94,7
	CV	0	1.292	100,0
	(%) Total	46,9	53,1	97,4

Tabela 2.4: Desempenho da classificação por ANN com MPL e usando banco de dados SMOTE.

		Preditos		
Subconjuntos	Nível da lesão	SV	CV	(%) Acerto
Treinamento	SV	138	5	96,5
	CV	20	96	82,8
	(%) Total	61,0	39,0	90,3
Teste	SV	53	4	93,0
	CV	5	32	86,5
	(%) Total	61,7	38,3	90,4

Fonte: Autora (2018).

A técnica *oversampling* resulta em uma melhora significativa da precisão geral de classificação da ordem de 97,4%, conforme ilustra a Tabela 2.3. Todavia, se verifica que o classificador foi tendencioso, resultando em falsos positivos para os acidentes CV. A técnica SMOTE, Tabela 2.4, por sua vez, evita e detecta *overfitting*, a partir do treinamento do classificador supervisionado considerando a distribuição dos dados originais e a análise de vizinhança dos dados, o que proporciona uma precisão geral de 90,4% e em uma taxa de acerto de 96,5% para os acidentes SV e de 82,8% para os acidentes CV.

A seguir apresenta-se a avaliação da curva ROC (*Receiver Operating Characteristic*), expressa pela relação entre a sensibilidade e a especificidade do modelo (Fawcett, 2006). A sensibilidade corresponde a proporção de verdadeiros positivos, ou seja, avalia a capacidade do modelo classificar corretamente uma observação. A especificidade equivale a proporção de verdadeiros negativos, isto é, a capacidade do modelo em prever observações incorretamente.

A Figura 2.9 (a) apresenta a curva ROC, na qual se verifica que a AUC assumiu valores médios da ordem de 0,641, para as variáveis alvo acidentes SV e CV, considerando o banco de dados desbalanceado. Ressalta-se que valores de AUC no intervalo de $0,7 \leq AUC < 0,8$ fornecem uma modelagem com discriminação aceitável, enquanto que valores entre $0,8 \leq AUC < 0,9$ expressam uma modelagem excelente e, por fim, valores

de $AUC \geq 0,9$ uma modelagem com elevado potencial de discriminação (Hosmer e Lemeshow, 2000).

A Figura 2.9 (b) apresenta a curva ROC para cada variável dependente (SV e CV) com a aplicação da técnica de amostragem *undersampling*. Ao analisar o gráfico da curva ROC (Figura 2.9 (b)), verifica-se que os valores preditos apresentaram AUC com valores médios de 0,895 tanto para os acidentes SV quanto para os acidentes CV. A modelagem da severidade dos acidentes (SV e CV) pela *oversampling* resultou em uma precisão geral com $AUC \geq 0,7$ conforme ilustra a Figura 2.9 (c). A Figura 2.9 (d) apresenta a curva ROC para cada variável dependente SV e CV em função da aplicação da técnica de amostragem SMOTE. Neste caso, a curva ROC apresenta valores médios iguais a 0,952 para as predições das classes de acidentes SV e CV.

Figura 2.9 – Desempenho da classificação com ANN utilizando MPL em bases de dados de acidentes utilizando técnicas de amostragem.

Legenda: classificação utilizando bases de dados desbalanceadas (a); classificação com dados amostrados pela *undersampling* (b); classificação com dados amostrados pela *oversampling* (c) e classificação com dados amostrados pela SMOTE (d).

Fonte: Autora (2018).

2.4 Redes Complexas

O conceito fundamental de uma rede considera uma coleção de vértices unidos por arestas. Cada um dos pontos recebe a denominação de vértices ou nós, enquanto que as linhas ou arestas são referidas como links (Newman, 2010). Desta forma, conexões ou relações podem ser estabelecidas de forma objetiva entre os diferentes objetos do mundo real, por meio da representação virtual em redes, o que facilita o estudo das propriedades, do comportamento, da estrutura e das interações dos elementos que compõem uma amostra ou objeto no âmbito das mais diversas ciências.

Uma vantagem da utilização de redes complexas refere-se ao fato de que os vértices e as arestas podem ser rotulados com informações adicionais, que permitem descrever com

mais detalhes as características do sistema modelado. Ademais, pode-se fazer previsões matemáticas sobre os processos que ocorrem em redes como, por exemplo, como se dará o fluxo de informações pela internet ou a forma como uma doença irá se propagar por uma comunidade (Newman, 2010).

Em virtude de sua generalidade, as redes podem ser aplicadas a quase todos os sistemas, pois existem diferentes ferramentas de medições ou cálculos que permitem solucionar e responder perguntas específicas e questionamentos de determinados problemas. Representam, portanto, um meio geral, porém poderoso, de representar padrões, conexões ou interações em sistemas naturais dinâmicos e complexos do mundo real, nos quais se tem uma diversidade de propriedades e características.

Embora as redes estejam presentes em muitas áreas do conhecimento, nem todas são similares. Existem tipos especiais de redes, como redes direcionadas, redes não direcionais, redes ponderadas e assim por diante. Embora as redes tenham definições diferentes, na maioria dos casos, podem ser aplicados métodos simples de abordagens de análise em redes.

2.4.1 Conceitos básicos

Formalmente, uma rede R pode ser descrita pela relação $R = (V, E)$ em que V corresponde ao conjunto de vértices, $V = \{v_1, v_2, \dots, v_n\}$, e E ao conjunto de arestas ou links, $E = \{e_1, e_2, \dots, e_n\}$, (Newman, 2004a).

Uma aresta é definida pela relação $e_{v,u} = \{(v, u) = (u, v) \mid v, u \in V\}$, em que cada aresta conecta no máximo dois vértices, v e u , denominados de extremidades da aresta $e_{v,u}$. Portanto, os vértices v e u são vizinhos ou adjacentes e pode ser denotado por $u \sim v$. O número de vizinhos do vértice u é denominado grau (*degree* - d) de u , expresso algebricamente por $d(u) = |\{v \in V \mid u \sim v\}|$, (Kunegis, 2014).

Na modelagem por redes complexas, algumas considerações devem ser realizadas acerca da forma com que os vértices e links armazenam as informações. Uma rede pode ser classificada como: multimodal, quando as informações estão presentes nos vértices; como multirrelacional ou multidimensional, quando as informações estão presentes nas ligações (Wasserman e Faust, 1994); ou como heterogênea, quando apresenta vértices de tipos diferentes (Han, 2009).

Quanto aos links as propriedades fundamentais referem-se ao direcionamento e a valoração (Newman, 2010; Soares e Prudêncio, 2012). O direcionamento indica o tipo de relacionamento entre os vértices que compõem a rede. Os relacionamentos podem ser direcionais dando origem as redes direcionais (orientadas ou dígrafos), em que para cada link um vértice atua como transmissor e o outro receptor. Quando a relação entre os vértices é recíproca, a rede é chamada de não direcionada ou não orientada, sendo os links representados graficamente por linhas sem a presença de direcionamento na conexão.

A valoração é a propriedade que atribui um peso discreto ou contínuo a uma dada relação. Quando os links de uma relação são valorados, o conjunto $w = \{w_1, w_2, \dots, w_n\}$, que corresponde ao vetor que contém os pesos de cada link é adicionado a rede. Desta forma, a rede passa a ser representada por $R = (V, E, W)$, (Newman, 2004a).

Outro aspecto relevante que deve ser considerado no estudo de redes complexas refere-se a estrutura topológica, que relaciona informações globais e locais, que podem ser avaliadas e quantificadas por meio de medidas específicas utilizadas em análises de redes. A estrutura topológica pode revelar características qualitativas implícitas da rede analisada.

A seção a seguir apresenta as principais métricas utilizadas para caracterizar vértices, arestas e, por conseguinte, a própria rede.

2.4.2 Medidas de centralidade

A propagação de informações entre os nós que constituem uma rede está diretamente relacionada à distância mais curta entre eles, ou seja, ao número de caminhos mais curtos que os conectam, bem como a taxa de transmissão que percorre tais conexões. Para tanto, se torna necessário o uso de métricas que possibilitem o cálculo destes valores para os mais diferentes casos.

As medidas de centralidade são propostas com a finalidade de revelar a importância de vértices e arestas em um dado contexto. Dentre as medidas de centralidade mais conhecidas se destacam a centralidade de grau (*Degree Centrality* - DC), a centralidade de intermediação (*Betweenness Centrality* - BC) e a centralidade de proximidade (*Closeness Centrality* - CC), (Bao *et al.*, 2017; Bonacich, 1987 e Newman, 2010).

A centralidade de grau é um índice direto, proposto por Bonacich (1987), que corresponde ao número de vértices diretamente ligado a um vértice específico. Representa o número de arestas incidentes em cada vértice, ou seja, o número de vértices diretamente

ligado a um vértice específico. No entanto, vértices mais periféricos em uma rede com uma topologia específica podem apresentar uma influência significativa global e uma ‘pequena contribuição local para a vizinhança. Portanto, a centralidade de grau pode não ser apropriada como uma medida de análise global de uma rede ou para problemas de grande escala.

Desta forma, dada uma rede não direcionada representada por $R = (V, E)$, com V vértices e E arestas, sua estrutura pode ser descrita por uma matriz de adjacência ou matriz de relacionamentos A , com n vértices e de dimensões $n \times n$. Denotada, formalmente, por $A = a_{i,j_{n \times n}}$, onde a refere-se aos elementos da matriz de adjacência e os índices i e j aos nós ou vértices pertencentes a V . Quando $a_{i,j} = 1$, o nó i está conectado ao nó j e quando $a_{i,j} = 0$, o contrário (Newman, 2004b).

A centralidade de grau (DC) do nó i é definida como o número de arestas incidentes (Newman, 2004b):

$$DC(i) = \sum_{j=1}^n a_{ij} \quad (27)$$

A *Betweenness*, proposta por Freeman (1977), é definida como a fração dos caminhos mais curtos ou caminhos geodésicos que conectam todos os pares de nós que passam por um determinado vértice. Ela corresponde à média dos menores caminhos que passam por um vértice, sendo a medida mais utilizada para explicar o fluxo de informações em redes viárias e demais estudos de transportes, pois indica os vértices que aparecem com maior frequência entre os caminhos ou links. Quanto maior o valor da *Betweenness*, maior a interação de um vértice com os seus vizinhos.

Desta forma, a *Betweenness* (BC) de um determinado vértice $v \in R$ se baseia na contagem dos menores caminhos calculada pela seguinte expressão (Brandes, 2001):

$$BC(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \quad (28)$$

onde σ_{st} compreende a quantidade de menores caminhos que ligam todos os pares de vértices s , v e $t \in V$, e $\sigma_{st}(v)$ é a quantidade de menores caminhos que ligam todos os pares de vértices que passam por v .

A *Betweenness* analisa a frequência com que um vértice aparece no caminho geodésico entre dois outros vértices quaisquer, ou seja, essa medida tenta identificar vértices com grande potencial de fluxo de informação. Em redes metroviárias, por exemplo, os vértices

com maiores valores de *Betweenness* são os que poderão realizar com maior rapidez o processo de deslocamento para diferentes regiões. Os vértices com valores numéricos pequenos de *Betweenness* possuem distância relativa pequena em relação aos demais vértices de uma comunidade. Por isso, tendem a pertencer a uma mesma comunidade e a receberem informações em uma mesma taxa de probabilidade ou transmissão através dos menores caminhos.

A *Closeness* (CC), proposta por Sabidussi (1966), é definida como a soma da distância mais curta (distância geodésica) de um determinado vértice para os demais vértices da rede. Corresponde ao inverso da distância média do caminho mais curto (Brandes, 2001; Newman, 2010):

$$CC(v) = \frac{1}{l_i} = \frac{n}{\sum_j d_{i,j}}, \quad (29)$$

onde l_i indica a distância média entre os vértices i e j ; n ao número de vértices da rede e $d_{i,j}$ corresponde ao comprimento do menor caminho entre os vértices i e j , ou seja, ao número de arestas detectadas ao longo do caminho. Portanto, os menores valores de CC indicam que os vértices estão em uma mesma vizinhança ou o quanto as observáveis são semelhantes entre si. A proximidade indica o quão próximo este vértice se encontra de todos os demais vértices da rede.

O cálculo das medidas de centralidade utilizadas neste trabalho foi realizado com o auxílio da biblioteca IGraph¹, disponível nas linguagens *R* e *Python*, que apresenta o estado da arte em algoritmos baseados em métricas para análises de grafos, redes complexas e estruturas de comunidades. Dentre as vantagens de utilização da IGraph se destaca a possibilidade de adaptação dos algoritmos de acordo com a estrutura dos dados da rede que se deseja modelar (Csárdi e Nepusz, 2006).

Na literatura podem ser encontradas inúmeras medidas de caracterização de redes, que podem ser analisadas com mais detalhes nos trabalhos desenvolvidos por Barabási (2002), Barabási *et al.* (2002), Costa *et al.* (2007), Dorogovtsev e Mendes (2003) e Newman (2003).

¹Biblioteca Igraph de medidas de centralidade disponível para *download* no endereço: < <http://igraph.org/redirect.html> >. Acesso: julho/2017.

2.4.3 Medidas de desempenho de agrupamentos em redes

As redes complexas foram construídas a partir da abordagem de construção de redes complexas homogêneas, pelo método dos k – vizinhos mais próximos, usando o algoritmo hierárquico divisivo k – NN (k – *Nearest Neighbors*), (Berton, 2016, 2017). Este algoritmo escolhe a cada iteração os k vizinhos mais próximos com base no maior número de atributos correlacionados.

Desta forma, o algoritmo k – NN impõe que um vértice se conecte aos seus k vizinhos mais próximos, mesmo que alguns deles estejam distantes ou apresentem alguma dissimilaridade de vizinhança. A Figura 2.10 ilustra a construção de redes pela abordagem dos vizinhos mais próximos.

Figura 2.10 – Construção das redes complexas pelo método dos vizinhos mais próximos.

Fonte: Autora (2018).

Para cada rede gerada aplicou-se o algoritmo de detecção de comunidades *Fast Greedy* (Clauset *et al.*, 2004), que agrupa variáveis (observações) automaticamente em uma comunidade de acordo com suas ligações (conexões) e pesos dados a estas conexões (como por exemplo, a distância euclidiana entre os vértices analisados). A identificação das comunidades é realizada com base em uma função que identifica sub-grafos densos em uma rede, utilizando a otimização da medida de modularidade, proposta originalmente por Newman (2004c), por meio de uma busca gulosa em um mapa hierárquico (dendograma) com a finalidade de encontrar a divisão da rede que maximiza a modularidade. A Figura 2.11 ilustra a detecção de comunidades em redes homogêneas, em que os vértices de diferentes cores pertencem a diferentes comunidades.

Figura 2.11 – Detecção de comunidades em redes complexas.

Fonte: Autora (2018).

O cálculo da modularidade é obtido de acordo com a equação (30), com a finalidade de maximizar a medida da modularidade (Clauset *et al.*, 2004).

$$Q = \frac{1}{2m} \sum_{i,j} (A_{i,j} - \frac{k_i k_j}{2m}) \delta(c_i, c_j), \quad (30)$$

em que m corresponde ao número total de arestas da rede e $A_{i,j}$ corresponde a matriz adjacência entre os vértices i e j (caso exista conexão entre os vértices i e j tem-se que $A_{i,j} = 1$, caso contrário $A_{i,j} = 0$), k_i e k_j correspondem, respectivamente, ao número de vértices conectados aos vértices i e j (ou seja, k_i equivale ao grau do vértice i e k_j ao grau do vértice j). A função δ retorna 1 caso os vértices i e j pertençam a mesma comunidade e 0 caso contrário. A Q assume valores entre 1 e -1. Na prática, considera-se uma estrutura válida de comunidades valores de Q entre 0,3 a 0,7, uma vez que raramente a modularidade assume valores próximos a 1. No entanto, quanto mais próximo a 1 maior será a probabilidade de existirem comunidades reais presentes na rede, ou seja, da rede apresentar uma estrutura robusta de comunidades.

Nesta pesquisa, a modularidade foi calculada para cada nível de agrupamento usando a biblioteca da IGraph². O máximo valor encontrado indica a melhor divisão encontrada pelo algoritmo para a rede analisada.

²Biblioteca de modularidade Igraph disponível para *download* no endereço: <http://igraph.org/python/doc/igraph-pysrc.html#Graph.modularity>. Acesso: agosto/2018.

2.4.4 Redes bipartidas

Muitas redes complexas podem ser descritas como redes bipartidas, que representam uma classe poderosa e generalizada de métodos de análise e modelagem no âmbito de redes, que com apenas algumas modificações podem ser utilizadas na identificação e visualização de comunidades, classificação e na predição de links.

Uma rede bipartida, também chamada de rede de dois modos, é uma rede cujos nós podem ser divididos em dois conjuntos distintos, de modo que todos os links conectem dois nós de tipos diferentes (Newman, 2010).

Uma rede bipartida B pode ser descrita formalmente por $B = (V_1, V_2, E)$, em que $V_1 = \{v_1, v_2, \dots, v_i\}$ e $V_2 = \{u_1, u_2, \dots, u_j\}$ são dois conjuntos de vértices disjuntos e E o conjunto de links. A Figura 2.12, ilustra uma rede bipartida:

Fonte: Autora (2018).

Neste tipo de rede, cada aresta conecta um vértice V_1 a um vértice V_2 , de tal modo que $V_1 \cup V_2 = V$. A relação $B = (V_1 \cup V_2, E)$ denota uma rede bipartida em que V_1 e V_2 tem, respectivamente, i e j vértices cujo somatório representa o total de vértices de B . Sendo que, as arestas $e_{m,n}$ existem somente entre vértices heterogêneos, ou seja, $e_{i,j} \in E (v_i \in V_1, v_j \in V_2)$, sendo E o número total de arestas presentes em B (Wang e Qin, 2016; Xu, Chen e Li, 2015).

A matriz adjacente A' de B é simétrica e tem a seguinte forma:

$$A' = \begin{bmatrix} 0 & A_{i,j} \\ A_{j,i}^T & 0 \end{bmatrix}, \quad (31)$$

onde $0_{i \times i}$ e $0_{j \times j}$ são matrizes de zeros, $A_{i \times j}$ e $A_{j \times i}^T$ são matrizes diferentes de zero, sendo que o índice T indica a matriz transposta de $A_{i \times j}$. Pode-se utilizar na prática a forma $A_{i \times j}$ para denotar a matriz adjacente de B , em que as linhas representam os vértices de V_1 e as colunas os vértices do conjunto V_2 (Wang e Qin, 2016; Xu, Chen e Li, 2015).

Formalmente, todas as redes $B = (V, E)$ são bipartidas quando existe uma bipartição de V de tal forma que as arestas em E conectam um vértice de uma partição com um vértice de outra partição (Kunegis, 2014). Isto decorre do fato de que a modelagem por redes bipartidas explica as principais propriedades de uma rede como agrupamento e distribuição de grau (Guillaume *et al.*, 2006; Guillaume e Latapy, 2004).

Toda rede bipartida pode ser tratada como uma rede unipartida, também denominadas de redes homogêneas, que apresentam somente um tipo de vértice, ignorando a partição de V . No entanto, nos últimos anos têm havido uma motivação crescente para analisar redes bipartidas, em particular, no que se refere à predição de links, detecção de estruturas de comunidade, investigação de sub-comunidades e classificação.

A vantagem de se utilizar redes bipartidas é que este tipo de modelagem não implica em grande perda de informação na grande maioria dos casos, uma vez que diversos sistemas do mundo real não contêm apenas um tipo de vértice. Em redes bipartidas podem existir vértices do mesmo tipo e vértices heterogêneos, sendo que as bordas só existem entre os vértices heterogêneos.

Na literatura, existem métodos de ponderação que permitem descrever de forma mais realista as relações topológicas em redes bipartidas homogêneas e heterogêneas, visando minimizar a perda de informações. Pode-se derivar, por exemplo, de uma rede bipartida duas redes homogêneas ponderadas ou não. Esse processo é denominado de projeção unipartida.

Nestes métodos, ao se fixar dois vértices pode-se analisar o número de vizinhos comuns entre estes, bem como a soma dos pesos das arestas entre os vértices e seus respectivos vizinhos em comum (Valejo *et al.*, 2017a). A Figura 2.13, ilustra a criação de uma aresta ponderada entre os vértices 4 e 5 com base em seus vizinhos comuns em uma projeção unipartida ponderada.

Fonte: Adaptado de Valejo *et al.* (2017a).

Na Figura 13, os vértices 4 e 5 possuem 3 vizinhos em comum $w(4,5) = |\{1,2,3\}|$, logo, o peso da aresta entre eles é 3 (Valejo *et al.*, 2017a).

A maioria dos sistemas complexos do mundo real compreendem redes heterogêneas, que apresentam vértices de mesmo tipo, vértices de tipos diferentes, arestas entre vértices de mesmo tipo (intra-camadas) e arestas entre vértices de tipos diferentes (inter-camadas), (Valejo *et al.*, 2017a).

A literatura apresenta, ainda, o conceito de redes multicamadas, que são redes que apresentam um número superior a dois conjuntos de vértices distintos. No entanto, neste trabalho, será explorado apenas o conceito de redes bipartidas heterogêneas com no máximo dois tipos de vértices, pois compreende um estudo introdutório deste tipo de modelagem em dados de segurança viária.

2.4.5 Predição de links

O desenvolvimento contínuo da teoria de redes complexas vem permitindo que muitos sistemas complexos da natureza e do mundo real sejam descritos utilizando a abordagem por redes (Barabási e Albert, 1999; Sun *et al.*, 2015). Um dos “*hot topics*” de pesquisa em redes complexas, a predição de links, que visa prever a probabilidade da existência de um link entre dois vértices em uma rede, tem sido explorada cada vez mais nos últimos anos pelas mais diversas ciências (Getoor e Diehl, 2005; Lü e Zhou, 2011).

Os links preditos podem contribuir para que novas interações em redes bioquímicas sejam descobertas ou que novas amizades sejam recomendadas em redes sociais. Muitos métodos de predição de links foram propostos com base em mecanismos de evolução. Fundamentalmente, as técnicas de predição de links se classificam em três categorias: a primeira compreende as técnicas baseadas no conceito de similaridade entre os nós, ou seja, quanto maior a similaridade entre determinados nós maior a probabilidade de haver

um link entre estes; a segunda é fundamentada em relacionamentos topológicos, em que se utiliza métricas baseadas nas informações da estrutura da rede para prever os vértices com maior probabilidade de link e, por fim, a terceira técnica baseia-se em conceitos oriundos das relações sociais na tentativa de prever links que descrevem o comportamento humano através dos vértices que compõem uma rede (Peng *et al.*, 2015).

Na literatura, existem métodos de predição de links baseados em similaridade locais e globais, métodos baseados na máxima verosimilhança, modelos probabilísticos e métodos híbridos (Lü e Zhou, 2011). Tratam-se de abordagens tradicionais, sendo que a mais difundida baseia-se em métodos de similaridade topológica, nas quais as informações topológicas são independentes do domínio de aplicação do problema, o que implica que esta abordagem possa ser utilizada para todos os tipos de redes (Soares e Prudêncio, 2012).

Desta forma, considerando o estado da arte em predição de links usando a abordagem de similaridades locais (Zhou, Lü e Zhang, 2009), tem-se que os índices de predição de melhor desempenho e resultado são: o índice dos vizinhos comuns (*Common Neighbors* – CN), proposto por Lorrain e White (1971), que conta o número de vizinhos comuns em uma dada vizinhança, o índice de Adamic-Adar (AA), proposto por Adamic e Adar (2003), que refina a CN atribuindo um maior peso aos vizinhos com menor quantidade de conexões e o índice de alocação de recursos (*Resource Allocation* – RA), apresentado Zhou, Lü e Zhang (2009), sendo semelhante ao índice AA porém com a adição da informação do grau do nó dos vizinhos mais próximos.

Nesta seção, são apresentados os algoritmos de vizinhança h – hop de um vértice u , formalmente definido por $\Gamma_h(u) = \{v \mid \text{há um caminho de comprimento } h \text{ entre } u \text{ e } v\}$. Logo, a vizinhança 1-hop, corresponde ao conjunto de vértices $\Gamma_1(u)$ adjacentes a u . corresponde na prática, ao algoritmo CN.

E, por sua vez, a vizinhança 2-hop, corresponde ao conjunto de vértices $\Gamma_2(u)$ distantes de u com comprimento $h = 2$, ou seja, 2-hops e assim sucessivamente (Valejo *et al.*, 2018).

Índice dos Vizinhos Comuns (CN - *Common Neighbors*)

O índice CN compreende uma métrica de predição de links introduzida por Newman (2001), que estabelece uma relação entre a quantidade de conexões comuns entre dois vértices quaisquer de uma rede e a probabilidade de haver uma conexão entre estes.

Desta forma, considerando os vértices x e y , o número de vizinhos comuns pode ser obtido a partir do número de vértices resultante da intersecção dos conjuntos $\Gamma(x)$ e $\Gamma(y)$ representado por:

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)|, \quad (32)$$

onde CN reflete a similaridade relativa entre os pares de vértices x e y em redes complexas homogêneas e " $|\cdot|$ " corresponde a cardinalidade de um conjunto.

Embora os índices baseados em CN sejam os mais tradicionais na predição de links, a aplicação dos mesmos se torna limitada em redes do mundo real, uma vez que em geral compreendem redes grandes e densas, com centenas a milhares de vértices e com estruturas topológicas heterogêneas. Nestes casos, a estrutura de predição de links requer o cálculo de semelhanças para um grande número de pares de vértices, o que implica em um aumento na complexidade de tempo computacional (Cui *et al.*, 2016). Ademais, o índice CN trata cada vizinho igualmente, ou seja, sem pontuar os pesos de cada link (Ding *et al.*, 2016).

2-Hop Common Neighbor

O algoritmo *2-Hop Common Neighbor* atribui para cada par de vértices x e y , pertencentes a uma rede não direcionada $R = (V, E)$, uma pontuação ($S_{x,y}$) de acordo com uma medida de similaridade. Quanto maior for o valor de $S_{x,y}$ maior será a semelhança entre x e y e, por conseguinte, quanto menor $S_{x,y}$ maior será a dissimilaridade entre x e y . Ademais, como R é uma rede não direcional tem-se que $S_{x,y} = S_{y,x}$, (Lü e Zhou, 2010).

Os links inexistentes ou preditos são classificados em ordem decrescente de acordo com suas pontuações ou escores. Portanto, os links que apresentam os maiores escores apresentam as maiores probabilidades de ocorrência. Para verificar a precisão do algoritmo o conjunto de links observados ou existentes (E) são divididos em duas partes: o conjunto de treinamento (E^T), tratado como o conjunto de observações conhecidas, e o conjunto de teste (E^P), que apresenta as observações preditas. Logo, $E = E^T \cup E^P$ e $E^T \cap E^P = \emptyset$, (Lü e Zhou, 2010).

Logo, o algoritmo *2-Hop Common Neighbors* ponderado, denotado por *WCN-Weighted Common Neighbors*, é obtido pela seguinte expressão:

$$S_{WCN}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(z, y)}{\log(1 + s(z))}, \quad (33)$$

onde z corresponde ao vértice que pertence a intersecção dos conjuntos $\Gamma(x)$ e $\Gamma(y)$, $w(x, z) = w(z, y)$ denotam o peso do link existente entre os vértices x e y . A variável S_{WCN} corresponde a força do link entre os vértices x e y , que emprega $\log(1 + s(z))$ para evitar valores negativos, pois $s(z)$ pode ser menor que 1 (Lü e Zhou, 2010).

2.4.6 Métricas de avaliação e validação

Para quantificar a precisão de diferentes métodos de predição de links, existem duas métricas padrão chamadas de área sob a curva AUC/ROC e precisão (Lü e Zhou, 2011; Herlocker *et al.*, 2004). Em geral, emprega-se o método AUC pois este permite avaliar a importância de um atributo específico considerando suas possíveis combinações com outros atributos. De acordo com Huang e Charles (2005), a AUC é a medida que apresenta a melhor acurácia na avaliação de diferentes algoritmos.

A estatística AUC pontua a proporção de características preditas, fornecendo um valor único de acurácia entre todos os limiares possíveis compreendidos entre 0 e 1, sendo que os valores próximos a 0,5 indicam ou sugerem baixa discriminação da modelagem e os valores iguais a 1 a perfeita discriminação e prevalência do modelo utilizado (Fielding e Bell, 1997; Lobo, Jiménez-Valverde e Real, 2008).

O valor AUC pode ser interpretado como a probabilidade que a pontuação atribuída a um link predito (não existente), escolhido aleatoriamente, seja maior do que a probabilidade de um link real (existente), também escolhido aleatoriamente. Desta forma, o valor de AUC pode ser calculado pela seguinte expressão:

$$AUC = \frac{n' + 0,5n''}{n}, \quad (34)$$

onde n corresponde as comparações independentes, n' a quantidade de comparações onde os links preditos têm maior pontuação e n'' a quantidade de comparações onde um link predito e um link real têm a mesma pontuação.

Caso todos os valores de similaridade forem gerados a partir de uma distribuição independente e identicamente distribuída (IID), a precisão deve ser de cerca de 0,5. Logo, quando o grau de precisão excede o valor 0,5 isso indica que o desempenho de um

algoritmo de predição de links é superior quando comparado aos demais algoritmos de PL aleatórios (Papadimitriou, Symeonidis e Manolopoulos, 2012).

A validação cruzada é amplamente utilizada em predição de links pois permite avaliar a capacidade de generalização de um modelo a partir de um conjunto de dados. Na prática, determina o quanto um modelo se ajusta a um novo conjunto de dados.

Neste trabalho utilizou-se a validação cruzada $k - fold$, em que um conjunto de dados é subdividido em k subconjuntos em k iterações. A cada iteração, um dos conjuntos de k é utilizado como conjunto teste (10%) e os outros subconjuntos $k - 1$ compõem o conjunto de treinamento (90%) pela estratégia de validação cruzada (*cross validation*). Desta forma, ao final do processo todos os dados serão testados (Kohavi, 1995; Lü e Zhou, 2010).

A precisão (P) da abordagem é obtida a partir o número de links preditos corretamente (verdadeiros positivos - V_p) e do número de links preditos incorretamente (falsos positivos - F_p), sendo expressa pela seguinte relação:

$$P(\%) = \frac{V_p}{V_p + F_p} \quad (35)$$

2.5 Trabalhos relacionados

Os trabalhos relacionados ao tema de pesquisa proposto foram selecionados a partir de uma Revisão Sistemática da Literatura (SLR - *Systematic Literature Review*). A SLR analisa e explora, de forma aprofundada, as lacunas e o vieses de pesquisa, com base em resultados teóricos e empíricos de um conjunto restrito de trabalhos (Kitchenham, 2004, Petersen *et al.*, 2008).

Na SLR, o problema de pesquisa é investigado em diferentes domínios de aplicação, por intermédio da revisão do estado da arte dos trabalhos já publicados sobre o assunto. Desta forma, possibilita que se identifiquem as principais tendências existentes na comunidade de pesquisa e que o pesquisador justifique e apoie seus argumentos na apresentação de uma contribuição original (Kitchenham *et al.*, 2009; Shadroo e Rahmani, 2018).

A metodologia de SLR aplicada neste trabalho é constituída por cinco fases, conforme adaptação do método proposto por (Fiorini e Jabbour, 2017; Jabbour, 2013): (i) planejamento, (ii) estratégia de pesquisa, (iii) desenvolvimento e classificação

taxonômica, (iv) visão geral da produção científica e (v) identificação das lacunas na literatura.

Planejamento

Inicialmente, foi elaborado um protocolo de pesquisa que contempla três etapas: a identificação dos objetivos e questões de pesquisa, a definição das estratégias de busca, e a determinação dos critérios de inclusão e exclusão dos estudos.

Objetivos e questões de pesquisa

Nesta pesquisa, a aplicação da SLR tem o objetivo de estudar a abordagem de agrupamentos, detecção de comunidades, classificação e predição em bancos de dados de acidentes rodoviários, a fim de identificar:

- 1) publicações que exploram técnicas de IA no estudo e mineração de bancos de dados de acidentalidade viária;
- 2) publicações cuja proposta principal se refere a métodos de agrupamentos e detecção de comunidades de acidentes em ambiente rodoviário;
- 3) publicações cuja proposta principal se refere a métodos de predição de acidentes em rodovias;
- 4) identificação das variáveis utilizadas, campo de aplicações e domínios explorados no contexto da Engenharia de Transportes e Geomática;
- 5) vantagens, limitações e desvantagens encontradas nos itens 1, 2, 3 e 4.

Para assegurar o cumprimento dos objetivos propostos, a principal questão de pesquisa que orientou este estudo foi: “*Quais são as atividades de pesquisa em Segurança Viária direcionadas ao estudo e prevenção de acidentes em rodovias, considerando diferentes possibilidades em Inteligência Artificial?*”.

Estratégias de busca

A seleção prévia dos trabalhos pesquisados considerou fundamentalmente três critérios: (i) seleção de bases de dados que fornecem acesso ao artigo na íntegra; (ii) seleção de trabalhos publicados no formato de artigo ou artigo de revisão, publicados exclusivamente em inglês, desde o início da base até o ano de 2017, e limitados às áreas de Engenharia, Matemática e Ciência da Computação; (iii) exclusão de bases de dados que indexam artigos proprietários ou que incluam resultados de artigos indexados em diferentes fontes como, por exemplo, o Google Acadêmico; e, finalmente, (iv) busca

restrita a termos da pesquisa detectados em títulos, resumos e palavras-chaves, de forma a assegurar que os termos selecionados sejam encontrados no artigo.

Com base nestes critérios foram selecionadas cinco bases de dados: Scopus, Science Direct, Web of Science, IEEE Xplore e Transportation Research Board (TRB). Embora cada base apresente seu próprio mecanismo de pesquisa na biblioteca digital, priorizou-se que os quatro critérios de busca fossem atingidos.

As seguintes expressões de pesquisa foram aplicadas nos campos título, resumo e palavras-chaves considerando os sinônimos dos termos em estudo, visando a construção do estado da arte na temática proposta e a identificação de tendências e perspectivas futuras:

Expressão de busca: (“road accident” or “traffic accident” or “road safety”) and highway and (urban or rural) and (“data mining” or “machine learning” or “artificial intelligence”) and (“principal component analysis” or “Decision Tree” or “rule extraction” or cluster or “neural network” or “Bayesian Network” or classification or prediction or mapping or “complex network” or “bipartite network” or “georeferenced network” or “community detection” or “Highway Safety Manual” or “empirical method of Bayes” or “before and after study” or “segments homogeneous” or “multivariate analysis” or “crash prediction models” or “crash severity” or “crash frequency”).

Em razão de cada base de dados apresentar uma sintaxe de busca específica, o que limita a quantidade de expressões regulares a serem inseridas no campo de pesquisa da biblioteca, há a necessidade de se readequar, sem a alteração semântica, a expressão de busca. Em determinados casos recomenda-se segmentar a expressão de busca e, posteriormente, realizar a interseção entre as mesmas para que desta forma não sejam detectados artigos duplicados em uma mesma base de dados. A detecção de artigos duplicados deve ocorrer apenas na integração entre os artigos coletados entre as diferentes bases analisadas.

Crítérios de seleção e extração de informações

Cada artigo retornado com base na estratégia de busca foi selecionado em função das informações contidas no título, resumo e palavras-chaves. Essa fase define ao pesquisador quais artigos devem ser lidos integralmente e quais artigos devem ser rejeitados de acordo com os seguintes requisitos: (i) artigos duplicados, ou seja, que pertencem a mais de uma base de pesquisa; (ii) artigos com acesso restrito ou que não foram adquiridos pelas instituições de ensino e pesquisa envolvidas no projeto; (iii) artigos que não estejam no

formato de busca pré-estabelecido dentre os quais teses, dissertações, trabalhos de conferência, capítulos de livros; (iv) publicações incompletas e que contenham apenas título e resumo; e, finalmente, (v) artigos escritos em outros idiomas que não o inglês.

Um total de 860 artigos, no período de início do repositório de cada base ao ano de 2017, foram encontrados usando a expressão de busca nas cinco bibliotecas digitais, conforme apresentado na Tabela 2.5.

Tabela 2.5: Resultado da pesquisa a partir da Expressão de busca.

Base	Total
Scopus	642
IEEE Xplorer	25
Science Direct	105
Web of Science	54
Transportation Research Board	34
Total	860

Entre os artigos encontrados, 67 foram identificados como duplicados e 559 foram rejeitados de acordo com os critérios de exclusão como título, resumo e não convergência com a temática proposta no que se refere a estudos de detecção de agrupamentos e comunidades, classificação e predição de acidentes. Após a fase de seleção foram aceitos 234 artigos para a fase de extração das informações, no período de 1999 a 2017, conforme apresentado na Tabela 2.6.

Tabela 2.6: Processo de seleção dos trabalhos considerando o título, o resumo, a leitura do artigo e a verificação dos critérios de inclusão e exclusão.

Tipo	Total
Duplicados	67
Rejeitados	559
Aceitos	234

Os 234 artigos aceitos tiveram seus títulos e resumos lidos com a finalidade de se extrair os metadados: título, resumo, palavras-chaves, autor, ano de publicação, periódico e filiação.

A distribuição das publicações por ano é apresentada na Figura 2.14, que mostra uma tendência crescente no número de publicações por ano.

Figura 2.14 - Evolução do número de publicações por ano obtidas a partir de cada uma das bases consultadas.

O pico de publicações foi observado no ano de 2016, com um total de 49 trabalhos publicados e que representa 21% do total de artigos analisados. O menor número de publicações ocorreu no período de 1999 a 2003 com um total de 32 trabalhos publicados e que representam 14% do total de trabalhos analisados.

Estratégia de pesquisa

Posteriormente, realizou-se a restrição dos resultados obtidos na fase (i) considerando o número de citações de cada um dos artigos no Google Acadêmico, o domínio de aplicação, as variáveis utilizadas, o cenário de estudo e o método utilizado (abordagem proposta, vantagens, desvantagens e limitações).

Obteve-se um total de 44 trabalhos que foram lidos integralmente, dos quais 56% foram publicados entre os anos de 2000 a 2014, conforme ilustra a Figura 2.15.

Figura 2.15 -Restrição do número de trabalhos consultados com base na estratégia de busca.

A filiação referente aos 44 artigos selecionados é apresentada na Figura 2.16. Os trabalhos analisados indicam que 34% das publicações pertencem aos Estados Unidos, 14% a China, apenas 2% ao Brasil e 50% aos demais países.

Figura 2.16 - Mapeamento da filiação dos autores cujos trabalhos foram selecionados.

A Figura 2.17 apresenta a nuvem de palavras obtida com base nos 44 trabalhos selecionados. Quanto maior o tamanho em que é representada a palavra maior é a frequência com que ocorre nos artigos selecionados nas cinco bases de periódicos pesquisadas.

Com base na Figura 2.17 verifica-se que as palavras mais recorrentes nos trabalhos analisados foram: Redes Neurais Artificiais (22%), Árvore de Decisão – CART (13%), Sistemas de Transportes Inteligentes (9%), Regras de Associação (7%), Algoritmos Genéticos (7%), Algoritmo de Florestas Aleatórias (7%), Máquinas de Vetores de Suporte (6%). Enfatiza-se, que a utilização de redes complexas homogêneas corresponde a apenas em 4% dos trabalhos analisados, enquanto que 25% abordam com menos frequência outros algoritmos.

Figura 2.17 - Nuvem de palavras obtidas a partir dos metadados dos 44 artigos selecionados.

A partir da métrica *Google Scholar Metrics* foram selecionados os 10 artigos mais citados, conforme descreve a Tabela 2.7.

Tabela 2.7: Os dez principais artigos selecionados utilizando a métrica “*Google Scholar*”.

Referência (Autor/Ano)	Título	Nº citações
Karlaftis e Vlahogianni (2011)	Statistical methods versus neural networks in transportation research: Differences, similarities and some insights	348
Mannering e Bhat (2014)	Analytic methods in accident research: Methodological frontier and future directions	333
Chang e Wang (2006)	Analysis of traffic injury severity: An application of non-parametric classification tree techniques	317
Anderson (2009)	Kernel density estimation and <i>k-means</i> clustering to profile road accident hotspots	308
Caliendo, Guida e Parisi (2007)	A crash-prediction model for multilane roads	264
Chang e Chen (2005)	Data mining of tree-based models to analyze freeway accident frequency	243
Marhavilas, Koulouriotis e Gemeni (2011)	Risk analysis and assessment methodologies in the work sites: On a review	243
Delen, Sharda e Bessonov (2006)	Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks	232
Thill (2000)	Geographic information systems for transportation in perspective	216
Chang (2005)	Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network	199

Desenvolvimento e classificação taxonômica

De acordo com a literatura analisada elaborou-se um sistema de classificação visando categorizar os 44 artigos selecionados. Esse sistema foi baseado em 9 dimensões que foram categorizadas utilizando-se uma combinação de números e letras, conforme ilustra a Tabela 2.8.

Tabela 2.8: Taxonomia das categorias por dimensão para análise dos artigos.

Dimensão	Categoria
1 - Contexto	A - País desenvolvido
	B - País em desenvolvimento
	C - Não se aplica
2 - Área de Concentração	T – Operações e monitoramento de tráfego
	I - Gestão de infraestrutura, manutenção e reabilitação
	P - Planejamento
	F - Operações aéreas, terrestres, transporte de carga
	E - Meio ambiente e Transporte
	S - Segurança viária
	Q - Transporte de passageiros
3 - Problema	PRE - Predição
	CLA - Classificação
	AGR - Agrupamento
	DET – Detecção de Comunidades
	VISIO – Visualização de Dados
4 - Foco	TET – Técnicas Estatísticas Tradicionais
	ANN – Redes Neurais Artificiais
	RL – Regressão Logística
	CART – Árvore de Decisão
	SIG – Sistemas de Informações Geográficas
	FDK – Função de Densidade de Kernel
	K – Clustering
	M – Análise de dados espaciais e/ou temporais
	PCA – Análise de Componentes Principais

Dimensão	Categoria
4 - Foco	CA – Análise de Correlação Canônica
	MLG – Modelos Lineares Generalizados
	RMN – Regressão Multinomial Negativa
	PO – Regressão de Poisson
	RBN – Regressão Binomial Negativa
	DEP – Problema de Decomposição
	HSM – Highway Safety Manual
	EB – Método Empírico de Bayes
	FLO - Algoritmo de Florestas Aleatórias
	GEN – Algoritmos Genéticos
	MB – Métodos Bayesianos
	RB – Redes Bayesianas
	SVM - Máquinas de Vetores de Suporte
	LCC – Agrupamento por Classes Latentes
	SIT – Sistemas Inteligentes de Transportes
	MUL – Análise de dados Multinível
	PROB – Modelo Probit
	NB – Naive Bayes
	ASS – Regras de Associação
	MMC – Cadeia de Markov-Monte Carlo
	BAL – Técnicas de balanceamento de dados
	MGL – Modelo Logit Multinomial
	NCC – <i>Nearest Neighbor Classification</i>
	SIMIO – Simulação de Tráfego
	NET – Redes Complexas Homogêneas
5 - Método de Pesquisa	QT - Quantitativo
	QA - Qualitativo
	CO - Conceitual
	RE - Revisão
	ES – Estudo de Caso

Dimensão	Categoria
6 - Benefícios de desempenho	A - Financeiro
	B - Ambiental
	C - Social
	D - Outros
7 - Número de Objetivos	A – até 1 objetivo
	B - até 2 objetivos
	C – 3 ou mais objetivos
8 - Posição no modelo analítico	A – Variável independente
	B – Variável dependente
	C – Variável intermediária
	D – Não se aplica
9 - Número de variáveis	A- Não se aplica
	B – até 5 variáveis
	C – até 10 variáveis
	D – até 15 variáveis
	E - até 20 variáveis
	F – até 25 variáveis
	G – até 30 variáveis

A Tabela A.1 (Apêndice A) apresenta os resultados da classificação de cada um dos 44 artigos analisados com base no sistema de taxonomia apresentado na Tabela 8.

A análise bibliométrica indicou que a maior dos artigos abordou a modelagem de dados de segurança viária em um contexto nacional, especialmente em países desenvolvidos (93%). Os trabalhos realizados em países em desenvolvimento tiveram um menor número de publicações (7%).

Em relação à área de concentração, Figura 2.18, se verifica que os trabalhos foram desenvolvidos predominantemente no âmbito da segurança viária (77%). Quanto ao problema de pesquisa verifica-se, na Figura 2.19, que a maioria dos trabalhos focou na classificação dos dados de acidentalidade viária (32%). Estudos direcionados exclusivamente a predição e ao agrupamento foram identificados respectivamente em 18% e 7% dos trabalhos analisados. Os demais trabalhos exploram as abordagens de agrupamento, classificação, detecção de comunidades e predição de forma combinada.

Figura 2.18. Área de concentração. **Figura 2.19.** Problema de pesquisa

Considerando os trabalhos selecionados, verifica-se que em 54% das publicações os autores exploraram técnicas de IA. As técnicas de IA mais amplamente utilizadas nestes trabalhos foram ANN (10%), clustering (8%) e CART (6%). Somente em 2% dos trabalhos os autores utilizaram a abordagem de NET. Os demais trabalhos exploraram a modelagem de dados de acidentes viários por métodos tradicionais como os TET (6%), RBN (6%), EB (4%) e HSM (3%).

Em relação ao método de pesquisa, Figura 2.20, verifica-se que predominantemente as publicações selecionadas retrataram pesquisas quantitativas, qualitativas e de estudo de caso (80%). Quanto aos benefícios de desempenho 80% consideraram, simultaneamente, o impacto financeiro, associados aos custos operacionais e a implantação de contramedidas de acidentes de trânsito, e social, conforme ilustra a Figura 2.21.

Figura 2.20 - Método de Pesquisa. **Figura 2.21 - Benefícios de desempenho.**

Quanto ao número de objetivos 95% dos trabalhos apresentam 3 ou mais objetivos (Figura 2.22). Em relação a posição da variável no modelo em 50% dos trabalhos a variável modelada é dependente (Figura 2.23). A Figura 2.24 revela que 36% deles utilizaram até 10 variáveis no processo de modelagem. Enquanto que 25% compreendem

a artigos de revisão ou que não se atém a este quesito e 14% dos trabalhos utilizaram até 20 variáveis.

Figura 2.22 - Número de Objetivos.

Figura 2.23 - Posição da variável no modelo.

Figura 2.24 - Número de variáveis utilizadas.

Em geral, estes trabalhos consideram as variáveis relacionadas a seguir:

- tipo de rodovia (urbana ou rural);
- pista: tipo de pavimento, condição da pista, presença de animais na rodovia, coeficiente de atrito do pavimento;
- variáveis geométricas: número de intersecções, número de curvas, raio médio de curvatura, comprimento da curva, deflexão da curva, comprimento da seção, número de faixas, largura média ou crítica das faixas, perfil da rodovia, inclinação do segmento, traçado da rodovia, sinalização (horizontal, vertical, obras), canteiro central, largura do canteiro central, acostamento, largura do acostamento, barreira ou defesa, número de rampas, tipo de rampa, gradiente médio, grau da curva horizontal, declividade longitudinal e greide vertical;
- parâmetros de calibração do HSM;
- fatores ambientais: condição do tempo, condição de visibilidade, condição de iluminação, precipitação anual, velocidade do vento, estação do ano, número de dias com precipitação e volume de precipitação anual;
- uso e ocupação do solo: tipo de área circunjacente (residencial, rural, comercial, etc.), número de paradas de ônibus, número de estacionamentos, zonas críticas, presença de entroncamento;
- características do condutor: severidade da lesão, idade, gênero, nível educacional, experiência de condução, perfil de personalidade, grupo da população, nível de renda, grau de exposição ao tráfego, distância percorrida, uso do cinto de segurança e condição de sobriedade;
- características dos acidentes: descrição do local do acidente, frequência de acidente, número de vítimas, causas prováveis, sentido da via, tipo de acidente, local do evento,

hora do dia, dia da semana, mês, número de veículos envolvidos, número de acidentes por quilômetro ou por segmento, número de ocupantes nos veículos, número de acidentes envolvendo ciclistas e/ou pedestres, número de estações de metrô, posição do passageiro e do motorista;

- tipo de veículo: passeio, passageiros ou de carga; valor, tipo e dano ao veículo;
- variáveis de tráfego: volume de tráfego, limite de velocidade, período de congestionamento, controladores de tráfego (polícia, sinalizações na via, sinalizações luminosas, radar, câmeras), velocidade média, tempo de congestionamento, fator de hora pico, percentual de caminhões, percentual de veículos de passeio, percentual de veículos de passageiro, número de faixas congestionadas.

A Tabela B.1 (Apêndice B) descreve as variáveis utilizadas em cada um dos trabalhos. A Tabela C.1 (Apêndice C) apresenta a abordagem utilizada em cada um dos trabalhos, bem como as vantagens, desvantagens e resultados alcançados.

O levantamento sistemático dos trabalhos analisados revela que as técnicas de IA mais amplamente utilizadas na modelagem de dados de acidentes rodoviários foram as ANNs (Huang *et al.*, 2016; Li *et al.*, 2008; Karlaftis; Vlahogianni, 2011; Mussone, Ferrari, Oneta, 1999). Das NBs (Chen *et al.*, 2016; De Oña *et al.*, 2013; Kwon, Rhee e Yoon, 2015; Xie, Lord e Zhang, 2007), das CARTs (Chang; Wang, 2006; Chen *et al.*, 2016; Kwon, Rhee e Yoon, 2015), das SVMs (Halim *et al.*, 2016; Li *et al.*, 2008), das FLO (Halim *et al.*, 2016) e dos algoritmos genéticos (Halim *et al.*, 2016; Teodorović, 2008).

Estas técnicas forneceram resultados compatíveis aos obtidos pelas técnicas estatísticas tradicionais, como a regressão logística multinomial (Halim *et al.*, 2016), a regressão binomial negativa (Ahmed *et al.*, 2011; Huang e Adel-Aty, 2010; Huang *et al.*, 2016; Karlaftis e Vlahogianni, 2011; Li *et al.*, 2008) e do HSM (Saha, Alluri e Gan, 2011; Young e Park, 2013). Contudo, a aplicação destas diferentes abordagens ainda se restringe ao tipo de dado, a dimensionalidade da base de dados e das características do problema abordado na modelagem (Kumar e Toshniwal, 2015).

Nesta perspectiva, alguns trabalhos abordaram a utilização de métodos de agrupamento visando reduzir a heterogeneidade da base de dados e a otimização do desempenho de algoritmos de classificação, no que concerne a redução do número de regras de decisão. Dentre estes trabalhos, se destaca o trabalho desenvolvido por Kumar e Toshniwal (2016), que exploraram os métodos de agrupamento por LCC e pelo *k-medóides* na extração de regras de associação. Destaca-se também o trabalho

desenvolvido por Depaire, Wets e Vanhoof (2008), que utilizaram a LCC na identificação de acidentes de tráfegos em segmentos homogêneos da rodovia.

Os autores Sohn e Lee (2003), realizaram a fusão entre diferentes técnicas IA como a PCA, clustering e ANN, com a finalidade de melhorar o processo de classificação da severidade dos acidentes. Alguns trabalhos destacam, ainda, o ganho do desempenho no processo de classificação a partir de conjunto de dados reamostrados, quando comparados à classificação aplicada diretamente em conjunto de dados originais (Park e Abdel-Aty, 2015; Nowakowska, 2017).

Os autores Lin, Wang e Sadek (2014b) exploram o potencial dos algoritmos de detecção de comunidades, com base na maximização da modularidade em redes complexas, na extração de regras de associação em bases de dados de acidentes rodoviários. Os resultados revelam que este algoritmo é eficiente na identificação dos fatores mais relevantes e conformação dos agrupamentos, contribuindo para a obtenção de regras de decisão que correlacionam esses fatores.

Outro trabalho de destaque foi desenvolvido por Ting, Bin e Qi (2016), onde se verifica a aplicação das redes complexas em dados obtidos em rodovias em tempo real, com a finalidade de gerar alertas de congestionamento a partir de comunidades de veículos. O método explorado analisa a tendência de crescimento do volume de tráfego, possibilitando identificar as faixas da rodovia mais propensas à congestionamento e emitindo alertas às concessionárias. Todavia, o trabalho se concentra de forma efetiva na detecção de comunidades.

No entanto, não foram identificados na literatura trabalhos que explorem a abordagem de redes complexas usando a modelagem de redes bipartidas na predição de acidentes. O que indica que pesquisas nesta área sejam relevantes ao entendimento mais amplo do fenômeno da acidentalidade viária, considerando a tendência de utilização de grandes bases de dados no processo de modelagem, o tratamento de problemas multidimensionais e abordagem direta ou indireta de conceitos relacionados às cidades inteligentes e sua interação com os sistemas de transporte.

2.6 Considerações Finais

Neste capítulo, inicialmente, foram discutidas as diferenças básicas existente entre a modelagem estatística tradicional e as técnicas de AM baseadas em IA, no tratamento de dados utilizados em transportes.

Posteriormente, foram introduzidos os conceitos básicos sobre técnicas de AM já consagradas no âmbito da segurança viária, como a PCA, clustering, a CART, a ANN e a BNN. O estudo se direcionou ao problema de redução de dimensionalidade, agrupamento e classificação de dados. Foram discutidos, também, técnicas de amostragem visando a otimização do desempenho dos algoritmos de classificação apresentados.

Em seguida, foi apresentado os conceitos e fundamentos envolvidos em redes complexas. A finalidade é apresentar a possibilidade de aplicação desta técnica na detecção de comunidades, destacando alguns trabalhos emergentes. Por fim, discutiu-se a possibilidade do desenvolvimento de modelagens de predição a partir da estrutura de redes complexas bipartidas, usando algoritmos baseados em vizinhança e a análise de links em redes.

3 METODOLOGIA PROPOSTA

Para avaliar a abordagem da pesquisa proposta foram utilizadas bases de dados de acidentes rodoviários georreferenciadas. A primeira base de dados compreende o histórico de acidentes registrados na Rodovia Dom Pedro I (SP-065), localizada no trecho urbano do Município de Campinas, no Estado de São Paulo, Brasil.

Os dados da SP-065 compreendem o histórico de acidentes ocorridos no período de 03/04/2009 a 29/08/2012 e de 30/08/2012 a 31/12/2016, que correspondem, respectivamente, aos períodos antes e depois da intervenção da concessionária “Rota das Bandeiras, no trecho urbano da rodovia compreendido entre o km 125 ao km 145,500. Trata-se de um trecho em pista dupla que vem passando, no presente momento, por obras de melhoria de infraestrutura e pela construção de segmentos marginais (Rota das Bandeiras, 2016).

A construção dos segmentos marginais foi dividida pela concessionária Rota das Bandeiras em 4 trechos. O “Trecho I - Marginal Norte”, do km 129 ao km 134, cuja obra foi iniciada em fevereiro de 2014 tendo já finalizado o trecho do km 129 ao km 133. O “Trecho I - Marginal Sul”, do km 129 ao km 131, cujas obras tiveram início em julho de 2016 e ainda se encontram em andamento. O “Trecho II”, do km 134 ao km 140, cujas obras tiveram início em agosto de 2012 e ainda se encontram em andamento. O “Trecho III”, do km 140 ao km 143, cujas obras tiveram início em 2014 e ainda se encontram em andamento. Por fim, o “Trecho IV”, que compreende o km 143 ao km 145 com início em abril de 2014 e cujas obras ainda se encontram em andamento.

O histórico de acidentes registrados na SP-065 é derivado do “relatório de atividades de acidente/recusa de atendimento” implantado pela concessionária Rota das Bandeiras. Pelo fato de apresentar informações confidenciais e restritas a concessionária e órgãos fiscalizadores, foram disponibilizadas para esta pesquisa informações de natureza pública, porém inéditas, uma vez que o desenvolvimento da pesquisa se deu paralelamente a coleta dos dados.

Os registros observacionais individuais de cada acidente apresentam a data da ocorrência, a hora, o marco quilométrico, a descrição da ocorrência (sem vítima, com vítima e com vítima fatal), o grau de severidade (vítimas ilesas, vítimas leves, vítimas moderadas, vítimas graves e vítimas fatais), o tipo de acidente (colisão traseira, colisão frontal, colisão transversal, colisão lateral, engavetamento, capotamento, atropelamento, tombamento, choque e outros), a causa provável (motorista, veículo, via/meio ambiente

e outros), o sentido da via (Norte e Sul) e o local do acidente (acostamento, faixa de rolamento, canteiro central, alça de entrada, alça de saída, faixa da direita e faixa da esquerda).

No entanto, foi verificado por meio da revisão sistemática da literatura e por meio de abordagens experimentais, que estas variáveis não atendem a metodologia proposta na pesquisa. Desta forma, foi solicitado à concessionária informações e dados que possibilitassem a reconstrução do cenário de cada acidente como estudos de melhoria de infraestrutura, tais como, ortofotos (antes e após a intervenção), dados derivados do georreferenciamento do trecho em formato KMZ, informações de tráfego (nível de serviço e volume médio diário anual de veículos de passeio e comerciais) e os projetos geométricos da rodovia antes e após a implantação das contramedidas de acidentes de tráfego.

Com base nos estudos preliminares fornecidos pela concessionária foi possível reconstruir o cenário de ocorrência de cada acidente, por meio da correlação espacial e temporal, e propor um protocolo de coleta de dados georreferenciados direcionados a estudos de agrupamento, detecção de comunidades e predição de acidentes em ambiente rodoviário. Estes dados foram explorados considerando uma abordagem diferente das observadas no processo de revisão sistemática da literatura, com a finalidade de contribuir com a transdisciplinaridade entre a Engenharia, a Matemática e a Ciência da Computação, contribuindo com um novo paradigma de análise de dados de acidentes rodoviários.

A abordagem proposta se baseia no estudo da potencialidade e das limitações das técnicas derivadas da IA e utilizadas na mineração de dados em processos de redução da dimensionalidade de bases coletadas sem delineamento experimental ou estatístico (PCA), no agrupamento e na classificação em estruturas de árvore (CART) e redes (neurais, bayesianas e complexas), com base em dados multiclases e desbalanceados, bem como no estudo da predição de acidentes fundamentada na teoria de redes georreferenciadas (bipartidas). Além disso, permite a discriminação dos dados de acidentalidade viária em termos de agrupamentos, comunidades ou predições realísticas no tempo e no espaço, bem como a flexibilidade da aplicação em estudos de diferentes escalas.

A pesquisa visa contribuir, de forma argumentativa e experimental, com a discussão da aplicação das técnicas tradicionais no domínio do espaço vetorial que, em geral, compreendem modelos aplicados a estudos em pequena escala e que apresentam baixa taxa de sucesso ao serem transferidos para o tratamento de problemas complexos e

multicausais, uma vez que se baseiam em dados populacionais e resultam nos melhores valores para uma determinada amostra, isto é, em valores médios.

Os métodos tradicionais são exclusivos para o tratamento de dados relacionais e que podem ser prontamente representados por redes. Todavia, os dados de segurança viária, compreendem uma base de dados proposicionais, representados no formato de Tabela atributo versus valor, o que justifica a utilização de redes bipartidas georreferenciadas, visando a minimização de perdas de informações na modelagem dos dados no formato de uma rede.

Assim sendo, são apresentadas nas seções a seguir o protocolo de coleta de dados com ênfase em segurança viária e o processo de modelagem de predição de acidentes utilizando a abordagem de redes complexas bipartidas. Por meio da predição de acidentes, usando a análise de links em redes, se identifica e analisa os pontos críticos ou com alta concentração de acidentes no trecho em análise, bem como o grau de severidade dos mesmos.

3.1 Protocolo de coleta de dados em ambiente rodoviário

As seções a seguir destacam as etapas necessárias para a coleta dos dados geoespaciais essenciais a modelagem de dados de acidentalidade viária. Estes dados serão integrados aos registros observacionais de cada acidente fornecido pela concessionária responsável pelo trecho em estudo.

3.1.1 Área de estudo

Conforme já citado, a área selecionada para a realização do experimento compreende um trecho urbano entre o km 125 ao km 145,500 da Rodovia Dom Pedro – I (SP-065), localizada no município de Campinas, no Estado de São Paulo, Brasil, conforme ilustra a Figura 3.1.

A SP-065 possui uma extensão total de 145,5 km e liga o Vale do Paraíba à região metropolitana de Campinas pela conexão com as seguintes rodovias: Rodovia Henrique Eroles, no município de Jacareí, Rodovia Fernão Dias, nas imediações do município de Atibaia e, finalmente, com a Rodovia Ademar de Barros, em Campinas, possibilitando o acesso ao Sul do Estado de Minas Gerais pela Rodovia Professor Zeferino Vaz, no município de Campinas. A rodovia se encerra no acesso ao km 104 da Rodovia Anhanguera.

Figura 3.1: Localização do Trecho em estudo – Escala: 1:500.000.

Fonte: Autora (2018).

O traçado da SP-065, atualmente concedida à concessionária “Rota das Bandeiras”, foi projetado com a finalidade de assegurar a ligação mais curta entre as rodovias Anhanguera e Dutra, sem a necessidade de cruzar os municípios que compõem a região metropolitana da cidade de São Paulo. A rodovia permite a ligação do fluxo de veículos que se desloca entre o interior paulista, o litoral norte do estado de São Paulo e o Rio de Janeiro, em conjunto com a Rodovia Anhanguera e o sistema Dutra-Carvalho Pinto. Além disso, ela possibilita o fluxo de veículos do interior do país à Belo Horizonte, pelo cruzamento com a rodovia Fernão Dias. Trata-se de uma rodovia com velocidade de projeto de 100 km/h, em grande parte de sua extensão, com exceção de determinados trechos próximos aos municípios de Nazaré Paulista e Itatiba.

3.1.2 Rastreio do trecho selecionado

Para a elaboração do protocolo de coleta de dados foram realizados rastreios por meio da tecnologia de posicionamento GNSS, nos modos relativo estático e cinemático, no sentido Norte e Sul da rodovia SP-065 no trecho compreendido entre o km 125 ao km 145,500, em dois dias consecutivos (14/12/2017 e 15/12/2017). Para tanto, foram utilizados dois receptores GNSS de dupla frequência da marca Leica Geosystems, modelo 1200 GX1230 GG.

Um dos receptores GNSS foi utilizado como estação de referência e foi instalado no Centro de Controle Operacional (CCO) da Concessionária Rota das Bandeiras, como ilustra a Figura 3.2. O outro receptor GNSS foi utilizado como remoto e permaneceu a bordo do veículo utilizado para percorrer o trecho em estudo. A antena do receptor móvel

foi instalada por meio de um suporte imantado na parte externa e superior do veículo, como ilustra a Figura 3.3.

Figura 3.2: Estação de referência no CCO da concessionária “Rota das Bandeiras”.

Figura 3.3: Antena do receptor GNSS remoto instalada na parte superior do veículo utilizado na coleta dos dados.

Fonte: Autora (2018).

Ambos os receptores coletaram dados simultaneamente com uma taxa de coleta de 0,5 segundos e com máscara de elevação de 0° , com a finalidade de registrar os dados de todos os satélites disponíveis no período de rastreo pelo método de posicionamento relativo cinemático. Portanto, a cada 0,5 segundos uma observação era registrada e com isso se obteve uma nuvem de pontos coletados em ambos os sentidos da rodovia.

O veículo percorreu o trecho com uma velocidade média entre 80 km/h a 90 km/h. Após a finalização do rastreo os pontos coletados foram processados no software Leica Geo Office (LGO). Nesta etapa, inicialmente, se fez o transporte de coordenadas, com a finalidade de se estimar as coordenadas de referência da base instalada no CCO, utilizando a técnica de posicionamento relativo estático, vinculando-as ao Sistema Geodésico Brasileiro (SGB), cujo referencial geocêntrico é o SIRGAS - Sistema de Referência Geocêntrico para as Américas (época 2000,4) por meio das estações da Rede Brasileira de Monitoramento Contínuo dos Sistemas GNSS (RBMC), (IBGE, 2018).

Para tanto, foram selecionadas as estações da RBMC próximas a região de rastreio compreendidas pelas estações SPC-1, instalada na Universidade Estadual de Campinas (UNICAMP), a SPBP localizada no município de Bragança Paulista e a POLI, instalada no campus da capital da Universidade de São Paulo.

Os dados das estações da RBMC encontram-se disponíveis em: <https://downloads.ibge.gov.br/downloads_geociencias.htm> (Acesso: julho/2018), no formato universal RINEX (*Receiver Independent Exchange Format*). A Tabela 3.1 apresenta as coordenadas geodésicas Latitude (φ), Longitude (λ) e Altitude Elipsoidal (h) homologadas para as estações da RBMC selecionadas no referencial SIRGAS 2000 (época 2000,4), com suas respectivas precisões ($\sigma_\varphi, \sigma_\lambda, \sigma_h$).

Tabela 3.1: Coordenadas geodésicas das estações da RBMC.

Coordenadas Geodésicas						
Estação	φ	σ_φ	λ	σ_λ	h (m)	σ_h
SPC1	- 22°48'58,63052"	0,001	- 47°03'45,69584"	0,002	622,980	0,008
SPBP	- 22°55'32,11795"	0,001	- 46°32'04,09220"	0,001	893,701	0,003
POLI	- 23°33'20,33228"	0,001	- 46°43'49,12319"	0,001	730,622	0,004

Fonte: IBGE (2018).

No processamento e transporte de coordenadas foram utilizadas órbitas precisas GPS e Glonass, disponibilizadas no endereço: <<http://www.gnsscalendar.com/>> (Acesso: julho/2018). Adotou-se o ângulo de corte de 15° e, portanto, todos os dados dos satélites coletados abaixo desta máscara de elevação foram excluídos visando minimizar os erros inerentes a propagação do sinal. Selecionou-se o modelo troposférico de Hopfield e a combinação linear *ion-free* visando, respectivamente, a minimização ou eliminação da refração troposférica e ionosférica. As Tabelas 3.2 e 3.3 apresentam, respectivamente, as coordenadas geodésicas e UTM (Universal Transversa de Mercator) calculadas para a estação de referência instalada no CCO:

Tabela 3.2: Coordenadas geodésicas das estações da RBMC.

Coordenadas Geodésicas						
Estação	φ	σ_φ	λ	σ_λ	h (m)	σ_h
Base	-22°57'05.03584	0,0075	-46°51'51,54246"	0,0077	714,9277	0,0204

Tabela 3.3: Coordenadas UTM estimadas para a base de referência do CCO.

Estação	E (m)	N (m)
Base	308.849,9432	746.0647,368

A partir das coordenadas de controle estimadas para a estação de referência realizou-se o processamento dos dados coletados em modo cinemático em ambos os sentidos da rodovia. Foram determinadas as coordenadas dos pontos pertencentes a nuvem de rastreo.

Foram coletadas as posições de passarelas, viadutos, barreiras de contenção, postes de iluminação e radar. As demais informações foram obtidas indiretamente e georreferenciadas por meio das coordenadas dos pontos pertencentes a nuvem de pontos como descrito a seguir.

Tipologia do trecho

Por meio das coordenadas dos pontos de rastreo (nuvem), se fez a correlação espacial entre o marco quilométrico e as informações tipológicas de cada trecho. Para tanto, foram utilizados os projetos geométricos da rodovia em formato CAD, no período antes-depois, para delimitar o início e o fim de cada segmento, bem como a visualização espacial destes segmentos no *Google Earth*. Desta forma, cada segmento foi discriminado como “reta” ou “curva”.

Obtenção da largura da pista, acostamento e canteiro central

Para coletar a largura da faixa de rolamento, do acostamento e do canteiro central, foram utilizados os programas AutoCad e o Google Earth, uma vez que não foi disponibilizado para o período “antes da concessão” o trecho com a Alça de Souza, compreendido entre os marcos quilômetros km 125 e km 128.

Para a extração da largura da faixa de rolamento, foram extraídas, no mínimo, três medidas para cada tipo de variável, para dados em formato CAD ou via o Google Earth. Posteriormente, adotou-se o valor médio destas medidas. Para a predição, considerando que o número de faixas varia de 3 a 4, para o segmento em análise, adotou-se o valor da faixa crítica em cada um dos segmentos.

Para obter a largura do canteiro central e do acostamento, foram adotados procedimentos análogos aos descritos anteriormente para a faixa de rolamento.

Passarelas, viadutos, barreiras de contenção, iluminação, radar e dispositivos de acesso

Em função da dificuldade de se coletar determinadas informações como iluminação, radar, defensas, passarelas, viadutos e dispositivos de acesso ou saída a rodovia (ramos, trevos e alça) com o rastreo GNSS cinemático, utilizou-se de forma complementar a ferramenta “*street view*” do *Google Earth*.

Desta forma, foi possível validar as informações obtidas em campo e correlacioná-las a cada sentido e segmento da rodovia.

Em relação as barreiras, registrou-se a gravação em vídeo do trecho percorrido, com a finalidade de identificar trechos sem barreiras, com barreiras e mistos, conforme ilustra a Figura 3.4.

Figura 3.4 - Padrões de barreiras de contenção.

Fonte: Autora (2018).

Sinalização horizontal, sinalização vertical e sinalização de obras

As sinalizações viárias da rodovia (horizontais, verticais e de obras) foram extraídas dos projetos em formato CAD fornecidos pela concessionária, por intermédio da ferramenta “*Google Stree View*”.

Em relação a sinalização vertical, a concessionária forneceu um inventário deste tipo de sinalização no período que antecede a implantação de obras. Desta forma, para o período de “depois” utilizou-se a ferramenta “*Google Stree View*”, para verificar os locais em que houve a alteração deste tipo de informação.

Velocidade máxima e radar

Para coletar a velocidade máxima de cada trecho utilizou-se a ferramenta “*Google Stree View*”. Cada quilômetro da rodovia foi percorrido utilizando-se essa ferramenta, com a finalidade de identificar o limite de velocidade presente nas sinalizações verticais,

verificadas *in loco* e no memorando de sinalizações verticais disponibilizado pela concessionária.

Dados de elevação

A elevação dos pontos foi obtida por intermédio do cálculo das coordenadas GNSS do rastreio realizado. Obteve-se, em geral, no sentido norte o RMS médio de 6 cm e no sentido sul de 5 cm.

Determinação do VDMA

O VDMA (Volume Diário Médio Anual) corresponde a quantidade de veículos que transitaram em um determinado trecho de uma rodovia ao longo do período de um ano.

Neste trabalho, a concessionária forneceu o VDM (Volume Diário Mensal), dividido em veículos de passeio e comercial, para cada segmento em análise. Desta forma, é necessário se realizar a conversão do VDM para VDMA, conforme recomendações do HSM (AASHTO, 2010). O VDM de veículos comerciais é obtido pela razão do VDM de veículos de passeio multiplicado por dois, já que um caminhão de eixos múltiplos apresenta comprimento significativamente maior que um carro de passeio.

Para a conversão correta dos valores de VDM para VDM_{ano} ou $VDMA$, utilizou-se a equação 37.

$$VDM_{ano} = \frac{VDM_{janeiro} \times \sum dias_{janeiro} + \dots + VDM_{dezembro} \times \sum dias_{dezembro}}{\sum dias_{meses\ utilizados}}. \quad (37)$$

A equação (37) realiza a conversão dos $VDMs$ em $VDMA$ s. Para tal, deve-se multiplicar o VDM do mês pelo número de dias de cada mês, repetindo este procedimento para todos os 12 meses e somando os seus valores. Posteriormente o valor calculado é dividido pelo somatório de dias dos meses utilizados.

O cálculo do VDM_{ano} é realizado para os veículos de passeio e veículos comerciais de forma independentemente. Posteriormente, se realiza o somatório dessas duas quantidades obtendo-se o VDM_{total} , como recomendado pelo HSM (AASHTO, 2010).

Os cálculos foram realizados tanto para veículos de passeio quanto para comercial e, posteriormente, a soma de ambos para se obter o valor final do VDMA. A Tabela 15 apresenta o VDM_{total} calculado para os anos de 2009 a 2016 (Apêndice D).

Para os períodos em que a concessionária não forneceu os valores de VDM_{ano} como, por exemplo, para os anos de 2009 e 2010, em razão do sensor que mede a quantidade de veículos não ter sido instalado; e para períodos em que foram detectadas falhas no sensor em alguns segmentos da rodovia, como nos anos de 2011 e 2016, realizou-se a estimativa do VDM_{ano} usando regressão linear e análise de tendência.

O banco de dados final da SP-065 passou então a ser composto por 43 variáveis, sendo elas: dia da semana (DY), mês (MM), ano (YYY), Severidade (SE), Tipo de Acidente (TA), Causa Provável (CP), período do dia (PE), marco quilométrico (KM), Sentido da via (SEN), Local do Acidente (LO), Vítimas Fatais (VF), Vítimas Não Fatais (VNF), Vítimas com danos Materiais (VM), Condição de Visibilidade (CV), Condição Especial (CE), Condição de Tempo (CT), Condição da Pista (CPI), Traçado da Pista (TP), Perfil da Pista (PR), Inclinação do Trecho (IT), acostamento (AC), largura do acostamento (LC), Luminosidade (LU), Radar (RA), Trevo (TR), Ramo (RM), Velocidade Máxima (VM), sinalização de obras (SO), Sinalização Horizontal (SH), Sinalização Vertical (SV), Latitude (LAT), Longitude (LON), Número de Faixas (NFX), Largura da Faixa Crítica (LFC), Canteiro Central (CC), Largura do Canteiro Central (LCC), Viadutos (VID), Pontes (POT), Barreiras (BAR), Volume Médio Diário Anual de Veículos de Passeio (VDMA_P), Volume Médio Diário Anual de Veículos Comerciais (VDMA_C), Volume Médio Diário Anual Total de Veículos (VDMA_T) e Intervenção (INTE). A descrição das instâncias das variáveis analisadas é apresentada na Tabela 16 do Apêndice E.

A Figura 3.5 ilustra o procedimento geral explorado para a obtenção da denominada “Base Doutorado”, que corresponde a integração entre os dados fornecidos pela concessionária e as informações geoespaciais obtidas direta e indiretamente.

A etapa inicial do trabalho consistiu na mineração de dados, que possibilitou a identificação de observações com ruídos e inconsistentes. A segunda etapa consistiu na reconstrução dos cenários dos acidentes pela correlação espacial e temporal dos registros observacionais e, por fim, os padrões de dados extraídos com base nas variáveis utilizadas no processo de modelagem de dados com ênfase em segurança viária.

Figura 3.5 - Processo geral de integração de registros observacionais de acidentes rodoviários com informações geoespaciais.

Fonte: Autora (2018). Será que não seria melhor aumentar os textos da figura?

3.2 Proposta de modelagem de dados usando técnicas de AM com ênfase na segurança viária

Nesta perspectiva, o conjunto de dados obtido por meio do protocolo de coleta proposto na seção 3.1, foi explorado com base em técnicas de AM fundamentadas em conceitos da IA.

Assim sendo, são apresentadas nas seções a seguir o processo de geração da base de predição com base em técnicas de AM como PCA, clustering e redes complexas homogêneas, a classificação da severidade dos acidentes de tráfego usando estruturas de dados em ANN e BNN. Em seguida, a modelagem de dados de segurança viária no formato de redes complexas bipartidas com dados no formato Tabela atributo-valor e, posteriormente, o processo de predição de links em redes, etapa que visa analisar se houve alterações no grau e modificação da acidentalidade viária por meio da predição dos tipos mais prováveis de acidentes que podem ocorrer em determinados quilômetros da rodovia a partir de atributos correlacionados.

3.2.1 Processo de geração da base de predição

Esta etapa visa obter uma nova base de dados denominada “*Base Prediction*”, contendo as variáveis de maior relevância ao estudo do fenômeno da acidentalidade viária considerando o trecho da rodovia em análise.

A Figura 3.6 apresenta, de forma geral, a metodologia proposta para a obtenção do conjunto de variáveis necessárias a predição de acidentes.

Figura 3.6 - Obtenção da “*Base Prediction*” por meio de técnicas de AM.

Fonte: Autora (2018). Ver tamanho dos textos.

Desta forma, a partir da “Base Doutorado” aplicou-se a técnica de PCA, por meio do aprendizado não supervisionado, com a finalidade de reduzir a dimensionalidade da base de dados e de se realizar a análise microscópica do trecho da rodovia em análise. Em seguida foram exploradas técnicas de agrupamento visando reduzir a heterogeneidade do conjunto de dados, em função da identificação das variáveis de maior contribuição para a ocorrência dos acidentes no trecho analisado.

Neste trabalho foi explorada a técnica baseada em clustering utilizando o algoritmo *k-means* e a técnica baseada em análise de vizinhança, usando o algoritmo de detecção de comunidades *Fast Greedy* (Clauset *et al.*, 2004), pela representação dos dados em redes complexas. As redes foram construídas pelo método dos *k-vizinhos* mais próximos, usando o algoritmo *k-NN* (*k-Nearest Neighbors*), (Berton, 2016, 2017).

As técnicas de agrupamento foram comparadas iterativamente, com a finalidade de obter a estrutura de agrupamento mais robusta para os dados avaliados. A partir dos agrupamentos obtidos foram extraídas regras de decisão usando o algoritmo de aprendizado supervisionado baseado em árvores (CART). O processo de obtenção de regras de decisão por meio da aplicação prévia de agrupamentos visa reduzir o número de regras geradas pela correlação das variáveis presentes no conjunto de dados. Desta forma, obtém-se um número reduzido de regras, porém, que apresentam a maior probabilidade de ocorrência. Ressalta-se que neste processo, o conjunto de dados deve

ser balanceado para que não ocorra sobre ajuste dos dados ao modelo. Neste trabalho, adotou-se a técnica de amostragem SMOTE (Fawcett, 2016), por ser aquela que se mostrou mais adequada a classificação usando bases de dados de acidentes rodoviários.

O percentual de acerto das CART possibilitou a investigação da técnica de agrupamento mais adequada para a detecção de agrupamentos de acidentes em ambiente rodoviário. Ao final deste processo é possível gerar um novo conjunto de dados, denominado “Base Prediction”, que contém as variáveis de maior peso ou maior contribuição estatística para o problema investigado.

3.2.2 Classificação da severidade dos acidentes rodoviários

Pelo fato da tomada da decisão pelas concessionárias e órgãos competentes não se restringir a um problema de decisão binária ou a regras do tipo “*se* → *então*” (CART), foram aplicados os métodos de extração de regras em redes usando a modelagem de ANN e BNN. A extração de regras em redes fornece recursos para responder a perguntas do tipo “Por quê?” e “Como?” feitas a partir do treinamento de um classificador.

As regras em redes podem ser utilizadas posteriormente na geração de um conhecimento baseado em sistemas inteligentes. Estas regras permitem que o especialista realize inferências lógicas de forma precisa acerca de um evento multicausal. Ao passo que as regras “*se* → *então*”, fornecem uma “solução fechada” por meio de uma estrutura hierárquica de decisão (Solange, 2003).

Neste estudo, a ANN foi construída com MLP, pelo treinamento supervisionado com base no algoritmo *back-propagation* (Facelli *et al.* 2011). Ao passo que na abordagem BNN foi utilizado o aprendizado supervisionado aplicando o classificador Naive Bayes. Ambos os classificadores produzem uma probabilidade ou pontuação para cada observação (instância), que representa na prática o grau ou probabilidade de uma observação pertencer a uma determinada classe ou nível de severidade. Por esta razão, estes classificadores são chamados de classificadores probabilísticos (Solange, 2003).

Todavia, tanto o algoritmo *back-propagation* quanto o Naive Bayes dependem do balanceamento das classes pertencentes ao conjunto de dados (Fawcett, 2016). Desta forma, neste caso, os dados devem necessariamente ser balanceados por meio de algum método de amostragem. Neste trabalho, o balanceamento dos dados foi realizado pelo método SMOTE (Fawcett, 2016), explorado em experimentos correlatos a esta pesquisa.

3.2.3 Modelagem da base de dados georreferenciada em uma rede bipartida

Um dos métodos mais populares de construção de redes é o k - NN (Berton, 2016, 2017). Este algoritmo permite conectar um vértice x_i , pertencente ao conjunto de dados $X = \{x_1, x_2, \dots, x_n\}$, aos seus k vizinhos mais próximos, com base em alguma medida de similaridade (geralmente, a distância euclidiana). Em outras palavras, a rede dos k -vizinhos mais próximos, para um conjunto de n objetos, em um espaço métrico X (conjunto de pontos no plano com distância euclidiana), compreende uma rede com X sendo o seu conjunto de vértices e uma aresta de u para v , de forma que a aresta $(e_{u,v})$ existe se v é um vizinho mais próximo de u , ou seja, se a distância de v para u não é maior do que de u para qualquer outro objeto de X .

Na literatura, são encontrados outros métodos de construção de redes, dentre os quais pode-se citar o *epsilon* vizinhança (Berton, 2016; Chang e Chen, 2005), no qual um vértice v se conecta com todos vértices u se a distância entre v e u é menor que " e ", ou seja, $\|(v,u)\| < e$. Outros métodos foram propostos por Berton (2016), tais como, "*rgcli*" e "*mutual knn*", e o método "*l2knn*", proposto por Anastasiu (2015).

No geral, as redes complexas são consideradas em um espaço abstrato, no qual a posição dos vértices não tem um significado particular. No caso de redes complexas homogêneas de interação, proteína-proteína ou redes bipartidas de atores e filmes, essa consideração é aceitável (Beckett, 2016). No entanto, existem redes onde a posição espacial ou geográfica dos vértices é particularmente importante, pois influencia a evolução da rede e suas características topológicas. Este é o caso das redes rodoviárias ou das redes de Internet, nas quais a posição das cidades ou dos roteadores podem ser localizadas em um mapa e as arestas correspondem a entidades físicas reais, como estradas e fibras ópticas (Robins, 2004). Este tipo de rede é denominado de rede geográfica, georreferenciada ou espacial e pode ser modelada usando estruturas bipartidas, nas quais as entidades físicas ou observações representam um tipo de vértice e suas localizações representam outro tipo de vértice. Neste caso, as arestas são incluídas entre as entidades físicas e suas respectivas localizações. Um exemplo didático é o caso de uma rede bipartida entre usuários e *check-in*, onde um dos tipos de vértices representa os usuários, o outro tipo representa os *check-in* e as arestas são incluídas entre os usuários e seus respectivos *check-in*. É importante ressaltar, como já descrito anteriormente, que em muitos casos, os conjuntos de dados com informações geográficas são naturalmente

relacionais (Krzanowski e Bina; 2010; Berton, 2016), como por exemplo, redes de energia, redes de aeroportos, redes de metrô e redes neurais (Zhang, 2008 e Spanurattana e Murata, 2011).

Neste trabalho, as bases de dados georreferenciadas contendo os registros observacionais integrados aos dados geoespaciais são armazenados em uma única tabela, visando correlacionar informações redundantes e/ou superabundantes. Tal característica possibilitou que os dados fossem modelados utilizando estruturas de dados em rede no formato Tabela atributo-valor. Neste formato não há a necessidade de se estabelecer pressupostos ou relações entre as observações (instâncias ou linhas da tabela), pela abordagem de redes complexas.

Neste contexto, como é de interesse da pesquisa entender as relações entre acidentes que estão de alguma forma ligados às suas localizações geográficas, necessitou-se explorar uma modelagem que permitisse que se modelasse a correlação entre os acidentes e suas respectivas localizações geográficas. Dentre as técnicas de AM exploradas neste trabalho, a modelagem de redes que mais se aproximou desse objetivo foram as redes complexas bipartidas.

A vantagem da modelagem de dados usando redes complexas bipartidas é que não há a necessidade de se explorar métodos de construção para a detecção de comunidades, como em redes complexas homogêneas. Neste caso, a rede é construída com base na real instanciamento das observações que compõem o banco de dados de acidentes rodoviários, o que implica em um considerável ganho computacional na manipulação dos dados.

Além disso, a arquitetura de dados por redes complexas bipartidas torna o processo de modelagem de dados mais tangível à área de engenharia e pesquisas desenvolvidas no âmbito da segurança viária, quando comparadas às ANN e BNN, que são consideradas “caixas pretas”, uma vez que não se pode estimular a priori o número ideal de camadas de dados, sendo este processo determinado de forma iterativa.

Com base nesta problemática, foi proposto um método de modelagem de dados de acidentes rodoviários com base em redes complexas bipartidas, assumindo-se que o conjunto de dados é IID e georreferenciado.

A Figura 3.7 ilustra a modelagem de predição proposta a partir de bases de dados de acidentes rodoviários utilizando a abordagem de redes complexas bipartidas.

Figura 3.7 - Proposta de modelagem de dados rodoviários com ênfase na predição de acidentes com base em redes complexas bipartidas. Não teria de indicar a fonte como nas demais figuras? Verificar tamanho dos textos.

O método proposto possui três etapas, que são: **(1)** criação do conjunto de vértices V_1 e V_2 , **(2)** criação das arestas e estrutura topológica inicial e **(3)** processamento da estrutura inicial, ponderação da rede e predição de links por meio de filtragem colaborativa.

(1) Criação dos conjuntos de vértices V_1 e V_2

A primeira etapa compreende a criação dos conjuntos de vértices e a instanciação da rede. De forma geral, um tipo de vértice irá representar as observações (registros de acidentes), e o outro tipo de vértice irá representar os marcos quilômetros em que estes acidentes ocorrem. Formalmente, considere-se o conjunto de acidentes $X = \{x_1, x_2, \dots, x_n\}$ com $x_i \in \mathbb{R}^m$, ou seja, $x_i = \{a_1, a_2, \dots, a_m\}$.

Esse conjunto de dados será então descrito como uma rede bipartida $G = (V, E, W)$, na qual $V = V_1 \cup V_2$, $V_1 \cap V_2 = \emptyset$ e V_1 é um conjunto de n vértices, no qual cada vértice $u \in V_1$ corresponde a um ponto $x_i \in X$.

Neste trabalho, considerou-se que uma dada variável a_i representa os marcos quilômetros (km) nos quais os acidentes x_i ocorreram. Sendo assim, selecionou-se todos os valores possíveis em a_i , sem repetição. Considerando o trecho em análise, os valores possíveis em a_i estão no intervalo $S = \{125, 126, \dots, 145, 145 + 500\}$ variando a cada 1 km, ou seja, S possui 22 valores possíveis. Portanto, V_2 será composto por um conjunto de

22 vértices e cada vértice $u \in V_2$ corresponde a um valor $j \in S$. A Tabela 3.4 ilustra um exemplo da base de dados de acidentes.

Tabela 3.4: Representação genérica da base de dados.

Acidentes	a_1	...	a_m	km
x_1	...	...	...	125
x_2	...	...	...	125
x_3	...	...	...	126
x_4	...	...	...	126
x_5	...	...	...	127
x_6	...	...	...	127
...	...	...	...	...
x_n	...	...	...	145 + 500 m

A Figura 3.8 ilustra a primeira etapa, que corresponde a instanciação da rede e criação dos conjuntos de vértices V_1 e V_2 com base na Tabela 3.4.

Figura 3.8 - Instanciação da rede e criação do conjunto de vértices V_1 e V_2 .

Fonte: Autora (2018).

Observa-se que nesta fase a rede ainda não possui arestas. Na próxima etapa será abordada a criação das arestas entre os conjuntos de vértices V_1 e V_2 .

(2). Criação das arestas e estrutura topológica inicial

Nessa etapa serão criadas as arestas entre os vértices do conjunto V_1 e os vértices do conjunto V_2 , gerando a topologia inicial e os conjuntos E e W . Neste caso, tem-se que

$E \subseteq V_1 \times V_2$. Assim, uma aresta entre $u \in V_1$ e $v \in V_2$ existe se o acidente u ocorreu no quilometro v . A Figura 3.9 ilustra a segunda etapa, a qual corresponde a criação das arestas entre os vértices (V_1 e V_2) e a geração da estrutura topológica inicial. Além disso, o peso de uma aresta (v,u) será inicialmente 1, porém, não se descarta a possibilidade de uma formulação na qual a aresta seja inicializada de maneira ponderada considerando alguma restrição ou característica.

Figura 3.9 - Etapa de criação das arestas e a topologia inicial da rede.

Fonte: Autora (2018).

Observa-se que a rede construída é desconexa e possui vários componentes, por exemplo, os vértices x_1 , x_2 e o km 125 representa um componente. Essa característica dificulta a aplicação de algoritmos de mineração e análise, por exemplo, algoritmos que utilizam informações de vizinhança, conceitos de inferência coletiva e principalmente propagação de informações tem seu desempenho prejudicado em redes desconexas. A redução do desempenho desse algoritmo está associada a perda do conceito de vizinhança, ocasionado pela obtenção de redes esparsas.

Ressalta-se que neste estudo não foi adotada a abordagem multinível direcionada a compactação da rede, ou seja, a obtenção de redes mais densas. Este procedimento permite selecionar a melhor estrutura da rede para realizar o processo de predição de links. Essa limitação é indicada como recomendação a trabalhos futuros.

3.2.4 Predição de links em redes complexas bipartidas

Os sistemas de recomendação são amplamente utilizados para sugerir produtos, serviços, conteúdos e itens de informações para usuários potenciais. Uma técnica comumente utilizada para esta tarefa é a filtragem colaborativa, a qual depende

unicamente dos dados de interação (transação, classificação ou associação) entre usuários e itens para fazer recomendações (ou previsão de interações).

Um algoritmo de filtragem colaborativa padrão, baseado em usuários, inicialmente deriva as vizinhanças dos usuários, identificando usuários semelhantes com base em suas interações sobrepostas ou classificações semelhantes de itens comuns. Em seguida faz recomendações com base nas experiências dos vizinhos de um usuário.

Comumente, as interações entre usuários e itens são modeladas como links (ou arestas) em uma rede bipartida que contém vértices representando usuários e itens. Assim, a estrutura bipartida captura informações sutis nas relações entre usuários e itens, conforme ilustra a Figura 3.10.

Figura 3.10 - Predição de links em redes complexas bipartidas.

Fonte: Autora (2018).

Sob esta representação de rede, o problema de recomendação pode ser visto como uma tarefa de seleção de links não observados para cada vértice e, portanto, pode ser modelado como um problema conhecido como predição de links. Observe, no entanto, que os dois problemas são diferentes: a recomendação visa tipicamente encontrar itens de interesse para usuários; a predição de links visa encontrar links que aparecerão no futuro.

Em termos gerais, a maioria das pesquisas de predição de links concentraram-se em redes unipartidas, como redes sociais. No entanto, o problema de recomendação também pode ser abstraído a partir de uma rede bipartida. Formalmente, considere uma rede bipartida $G = (V, E, W)$ e dado que V é dividido em dois conjuntos V_1 e V_2 . Neste caso, U denota o conjunto universal que contém todas as arestas possíveis $|V_1 \times V_2|$ entre vértices de tipos diferentes. Assim sendo, $E \subseteq U$ e $P = U - E$ representará o conjunto de arestas não existentes na rede. A tarefa de PL em redes bipartidas é descobrir links no conjunto P que possa ser observado no futuro na rede.

Na abordagem proposta, tem-se uma rede bipartida georreferenciada entre acidentes e quilômetros, onde uma aresta existe entre um acidente e seu respectivo quilômetro de ocorrência. O conjunto E representa a acidentalidade real em relação a cada quilometro, e o conjunto U representará o conjunto total e possível de acidentalidade em cada quilômetro. Assim, o conjunto de interesse $P = U - E$, irá representar o conjunto total de acidentalidade possível de previsão. Cabe ressaltar, ainda, que não se está interessado no conjunto P , mas apenas nos itens mais promissores desse conjunto. Cada aresta do conjunto de links preditos P terá uma pontuação ou score, o qual representa a intensidade do link ocorrer e é utilizada para ranquear os links de modo que links mais “promissores” possuem uma maior pontuação.

A Figura 3.11 ilustra um exemplo dos conjuntos de links mais promissores. As arestas cheias representam o conjunto E , as arestas tracejadas representam o conjunto P e a soma de todas as arestas representam o conjunto U . Cada um dos links do conjunto P irá receber uma pontuação e os links mais promissores serão selecionados.

Figura 3.11 - Ilustração de um conjunto de links promissores.

Por fim, somente alguns links, os mais promissores, irão compor as observações futuras. No caso da modelagem proposta neste projeto, cada link previsto representa um possível aumento da acidentalidade com características específicas em um determinado quilômetro. Ou seja, a modelagem não prevê um acidente na sua forma primária ou unitária. Por exemplo, a modelagem não irá dizer se ocorrerá ou não um acidente específico com vítima fatal no km 145.

De outra forma, a modelagem proposta informa que acidentes com características correlacionadas tem uma grande chance de ocorrer em um determinado quilômetro, ou seja, não é previsto um acidente unitário, é previsto características de acidentalidade que

irão ocorrer com maior frequência em determinados trechos da rodovia. Isso pode indicar, por exemplo, não só um novo tipo de acidentalidade em um determinado trecho, mas também pode indicar que um tipo de acidentalidade com características específicas já existente em um determinado quilômetro irá ocorrer com mais frequência, ou irá continuar ocorrendo.

Estes resultados podem servir como base para tomada de decisão em relação as contramedidas de acidentes de trânsito, que possuem determinadas características em trechos específicos.

Neste projeto, inicialmente, afim de corroborar com a hipótese proposta, se utilizaram os métodos e medidas mais tradicionais da literatura para predição de links. Inicialmente, se avaliará a filtragem colaborativa por meio da medida de similaridade 2-hops CN, discutido na Seção 2.4.8.

3.3 Considerações finais

Neste capítulo foi apresentada a metodologia proposta na modelagem de banco de dados de acidentes rodoviários por meio de técnicas de AM derivadas de conceitos da IA. Para tanto, inicialmente foi proposto um protocolo de coleta de dados em ambiente rodoviário, visando a detecção dos fatores chaves na ocorrência dos acidentes de tráfego, por meio de técnicas de redução da dimensionalidade (PCA), técnicas de agrupamento baseadas em clustering (*k-means*) e vizinhança (redes complexas).

Posteriormente, a classificação da severidade da ocorrência dos acidentes foi explorada por meio de modelagens de dados em árvore (CART) e redes (ANN e BNN), com a finalidade de correlacionar o maior número possível de variáveis e de apresentar o maior poder de generalização considerando diferentes conjuntos de dados (WI e NI).

Por fim, é apresentada uma abordagem de predição de acidentes por meio da análise de links em redes complexas bipartidas, utilizando algoritmos de filtragem colaborativa baseados em médias de similaridade e critérios de vizinhança em redes.

4 EXPERIMENTOS E RESULTADOS

Nesta seção serão abordados os principais experimentos realizados, bem como os resultados obtidos com o desenvolvimento da pesquisa considerando a metodologia proposta.

Na seção 4.1 são apresentados os resultados obtidos com a aplicação da técnica de PCA na base de dados de acidentes da rodovia SP-065 para o período de 2009 a 2016. Na seção 4.2 e 4.3 são apresentados, respectivamente, os agrupamentos obtidos a partir de acidentes com vítimas no trânsito considerando a variação do grau de severidade (danos materiais, fatal e não fatal), com base na análise de clustering e redes complexas homogêneas, explorando os bancos de dados de acidentes antes (2009-2012) e depois (2012-2016) da intervenção da concessionária “Rota das Bandeiras” no trecho em estudo.

Na seção 4.4, são apresentadas as regras de decisão em bases de dados de acidentes rodoviários, obtidas por meio dos agrupamentos detectados nas seções 4.2 e 4.3, com a finalidade de avaliar qual o método de agrupamento apresenta o melhor desempenho em termos de redução da heterogeneidade dos dados e na identificação das variáveis ou das características discriminatórias na formação de cada agrupamento.

A extração de regras de decisão permite revelar uma visão geral sobre as características das ocorrências envolvendo VM, VNF e VF, bem como identificar ou sugerir possíveis relacionamentos subjetivos entre o conjunto de dados analisados. A seção 4.5 discute a exclusão das variáveis pelo processo de agregação ou eliminação, em função do grau de relevância, e do contexto da temática abordada, a partir das análises obtidas na seção 4.4.

Na seção 4.6 é apresentada uma discussão do processo de classificação de dados em redes. Nesta fase, a partir dos resultados obtidos pelo algoritmo que apresentou melhor performance em relação as diferentes bases de dados, foram aplicadas regras de classificação usando a abordagem de ANN com MPL e BNN. Nesta etapa, os acidentes foram classificados em NI e WI, conforme recomendação do HSM (AASHTO, 2010). Em todos os processos de classificação considerou-se as bases de dados desbalanceadas e amostradas, com a finalidade de identificar as vantagens e desvantagens de cada uma das técnicas, bem como as limitações inerentes as bases de dados de segurança viária.

E, por fim, na seção 4.7 são apresentados os resultados obtidos com base na predição de acidentes utilizando a modelagem de redes complexas bipartidas, visando identificar pontos críticos de acidentes, bem como a variação ou modificação do grau de accidentalidade viária em cada km do trecho em estudo.

4.1 Análise da viabilidade da aplicação da técnica de PCA em dados multivariados aplicados à acidentalidade viária

Para exemplificar a aplicação da técnica de PCA explorou-se o banco de dados de acidentes rodoviários da SP-065 (2009-2016) obtido com o protocolo de coleta proposto, constituído por 43 variáveis (Apêndice E) com um total de 5.202 acidentes, dos quais 2.701 ocorreram no cenário antes da intervenção e 2.501 no cenário depois da intervenção. Estes acidentes resultaram em diferentes níveis de lesão. Todas as análises da técnica de PCA foram feitas por meio de rotinas computacionais implementadas no SPSS - *Statistical Package for Social Science for Windows*, versão 22, que é um *software* de análise estatística de dados, que permite realizar cálculos complexos e visualizar seus resultados de forma simples e autoexplicativa.

4.1.1 Redução da dimensionalidade do banco de dados e identificação das variáveis discriminantes

A aplicação da técnica de PCA resultou na redução do banco de dados de 43 variáveis originais (Apêndice E) para 34 variáveis, conforme ilustra a Figura 4.1, considerando as variáveis necessárias a exploração da metodologia proposta na pesquisa e as magnitudes das comunalidades, sendo estas variáveis: Tipo de Acidente (TA), Causa Provável (CP), Período do dia (PE), marco quilométrico (KM), Sentido da via (SEM), Local do acidente (LO), Vítimas Fatais (VF), Vítimas com Danos Materiais (VM), Vítimas Não Fatais (VNF), Condição de Visibilidade (CV), Condição do Tempo (CT), Condição da Pista (CPI), Traçado da Pista (TP), Acostamento (AC), Largura do Acostamento (LC), Luminosidade (LU), Radar (RA), Trevo (TR), Ramo (RM), Velocidade Máxima (VM), Perfil da Pista (PR), Inclinação do Trecho (IT), Sinalização de Obras na Pista (SO), Sinalização Horizontal (SH), Sinalização Vertical (SV), Latitude (LAT), Longitude (LON), Número de Faixas (NFX), Largura da Faixa Crítica (LFC), Canteiro Central (CC), Largura do Canteiro Central (LCC), Barreiras (BAR), Volume Diário Médio Anual de Veículos de Passeio (VMDA_P) e Volume Médio Diário Anual de veículos Comerciais (VDMA_C).

Figura 4.1 - Variação do percentual da variância acumulada em função da redução do número de variáveis do banco de dados pela técnica de PCA.

A Tabela 4.1 apresenta a variância total explicada através do critério da raiz latente a partir das 34 variáveis selecionadas, obtendo-se 13 componentes com autovalores maiores ou iguais a 1, ou seja, considerados significantes. Estes componentes apresentaram uma variância acumulada de 73,11%, valor este dentro do recomendado pela literatura, que é de 70% (Solanas, Manolov e Leiva, 2011).

Tabela 4.1 - Variância total explicada.

Componente	Autovalores			Variância Acumulada		
	Total	% de variância	% Acumulada	Total	% de variância	% Acumulada
1	4,03	11,86	11,86	4,03	11,86	11,86
2	3,36	9,87	21,73	3,36	9,87	21,73
3	2,56	7,52	29,25	2,56	7,52	29,25
4	2,02	5,94	35,19	2,02	5,94	35,19
5	1,93	5,68	40,86	1,93	5,68	40,86
6	1,80	5,29	46,16	1,80	5,29	46,16
7	1,61	4,74	50,90	1,61	4,74	50,90
8	1,54	4,53	55,42	1,54	4,53	55,42
9	1,45	4,25	59,67	1,45	4,25	59,67
10	1,40	4,12	63,80	1,40	4,12	63,80
11	1,08	3,19	66,99	1,08	3,19	66,99
12	1,07	3,15	70,14	1,07	3,15	70,14
13	1,01	2,97	73,11	1,01	2,97	73,11
14	1,00	2,94	76,05			
15	0,94	2,77	78,82			
16	0,87	2,56	81,37			
17	0,81	2,37	83,74			
18	0,74	2,18	85,93			
19	0,64	1,89	87,81			
20	0,57	1,67	89,49			
21	0,55	1,61	91,10			
22	0,49	1,44	92,54			
23	0,39	1,13	93,68			
24	0,36	1,05	94,73			
25	0,34	0,99	95,72			
26	0,26	0,76	96,48			
27	0,23	0,68	97,16			
28	0,23	0,67	97,83			
29	0,22	0,64	98,48			
30	0,13	0,37	98,85			
31	0,12	0,35	99,19			
32	0,11	0,32	99,51			
33	0,10	0,28	99,79			
34	0,07	0,21	100,00			

A Figura 4.2 apresenta em ordem crescente os valores das comunalidades obtidos para as 34 variáveis seleccionadas. Observando a Figura 4.2, verifica-se que a aplicação da técnica de PCA resultou em uma comunalidade média de 0,73 e em magnitudes mínima e máxima de, respectivamente, 0,42 e 0,92. As variáveis com maior comunalidade apresentam a maior capacidade de explicação em cada componente como, por exemplo, VDMA_C e VDMA_P com valores de, respectivamente, 0,91 e 0,92.

Figura 4.2 - Comunalidades das variáveis resultantes da PCA.

A análise do gráfico da Figura 4.3 indica 13 componentes, em que se deve observar o ponto no qual a curva desacelera, formando um “cotovelo” ou “joelho”. Ao utilizar o gráfico qualitativamente, o objetivo é localizar um ponto que separa uma parte de forte inclinação para a esquerda, de outra parte com inclinação mais rasa para a direita. O número de componentes principais em que a separação ocorre é então considerado como o corte de truncamento, o critério de inclinação não envolve inferência estatística quantitativa (Wilks, 2006).

Figura 4.3 - Gráfico com o número de componentes.

4.2 Clustering aplicado à detecção de agrupamentos de vítimas no trânsito

Neste trabalho optou-se por utilizar a análise não hierárquica pelo método k – means ao invés da análise hierárquica, pois este método é o mais utilizado na área de transportes

e se aplica a base de dados mistas, ou seja, que apresenta dados escalares e categóricos. O método k – means possibilita medir a importância interna e relativa das variáveis em cada cluster.

Todas as análises de cluster foram feitas no software SPSS considerando as variáveis selecionadas na seção 4.1. Para que a análise não sofra a influência das escalas de mensuração, as variáveis tiveram seus valores padronizados.

Os acidentes ocorridos entre os anos de 2009 a 2012 resultaram em 3.087 vítimas no trânsito, das quais 79% correspondem às VM, 20% as VNF e 1% as VF. Ao passo que para os anos de 2012 a 2016 foram registradas 2.891 vítimas no trânsito, das quais 77% correspondem às VM, 22% as VNF e 1% as VF. Estes dados foram utilizados apenas na etapa exploratória da pesquisa, com a finalidade de auxiliar na interpretação descritiva das bases de dados, com a redução da heterogeneidade dos dados, e na detecção de possíveis relações subjetivas entre as variáveis.

Os agrupamentos aqui denominados de clusters foram avaliados considerando independentemente o grau de severidade dos acidentes (VM, VNF e VF) e os cenários de intervenção antes e depois. Portanto, nesta etapa para cada cenário e grau de severidade foram obtidos 3 bancos de dados com 31 variáveis cada a partir da seção 4.1, visando exclusivamente a detecção de possíveis agrupamentos em registros contendo os diferentes níveis de severidade das vítimas no trânsito. Ressalta-se, que nesta abordagem exploratória, uma instância ou observação presente na base de dados geral de acidentes rodoviários (seção 4.1) pode estar presente em qualquer base de dados de severidade (VM, VNF e VF), uma vez que um registro de acidente pode ocasionar múltiplas lesões no trânsito.

Foram avaliados pelo método k – means agrupamentos de 3 a 10 clusters (k). Experiências anteriores realizadas com o conjunto de dados revelaram que para $k > 10$ não há variações significativas na formação dos agrupamentos em relação ao WSS, ou seja, em relação ao somatório dos erros quadráticos das instâncias de cada cluster. O WSS mede o grau de compactação de um agrupamento. Quanto menor o WSS maior será a homogeneidade dentro de cada agrupamento. Deve-se adotar um k a partir do qual não se verifica o aumento significativo no somatório total da variância dos dados com a adição de novos clusters.

Os clusters obtidos pelo método k – means foram avaliados em função do WSS e utilizando a métrica de desempenho de particionamento Silhueta. O WSS e o valor médio

da Silhueta foram analisados para diferentes valores de k , com a finalidade de encontrar o melhor valor para este parâmetro. Em geral, quanto maior o valor de k menor será o valor médio da Silhueta, ou seja, mais distante um ponto pertencente a um cluster estará em relação aos clusters vizinhos. Os valores de Silhueta próximos a 1 indicam que os clusters estão bem agrupados.

4.2.1 Detecção de clusters de acidentes de trânsito com vítimas com danos materiais utilizando cenários antes e depois da intervenção

A Figuras 4.4 (a) e 4.4 (b) apresentam, respectivamente, o WSS e a média do somatório dos desvios quadrados em relação ao número de clusters, para os acidentes com VM no período antes da intervenção (2009-2012).

Figura 4.4 - Clustering para a base com VM para 2009-2012.

(a) WSS para diferentes valores de k com VM para 2009-2012.

(b) Média do WSS para diferentes valores de k com VM para 2009-2012.

(c) Valor médio da Silhueta para diferentes valores de k com VM para 2009-2012.

(d) Frequência das observações com $k = 4$ para VM de 2009-2012.

O somatório da variância dos dados fornecido pelo WSS em relação ao número de clusters, Figura 4.4 (a), e o seu valor médio, Figura 4.4 (b), permitem verificar até que ponto o particionamento dos dados em clusters resulta em uma variação significativa dos agrupamentos obtidos. Para o banco de dados analisado (2009-2012) observa-se que para

$k > 4$ o WSS se estabiliza, apresentando uma discrepância não tão expressiva em termos de variância. Portanto, o “cotovelo” ou “joelho” do gráfico ocorre para $k = 4$, o que sugere que este seja o melhor valor para o parâmetro k .

Ao se analisar o valor médio da Silhueta para diferentes valores de k , verifica-se que para $k = 4$ o valor médio da Silhueta foi de 0,91, Figura 4.4 (c), o que sugere que os clusters encontrados apresentam uma estrutura robusta de agrupamento (Kaufman e Rousseeuw, 1990). Na Figura 4.4 (d) verifica-se que do total de 2.433 registros de acidentes envolvendo VM no período de antes (2009 a 2012), 517 ocorrências (21%) compõem o cluster 1, 753 ocorrências (31%) o cluster 2, 1.049 ocorrências (47%) o cluster 3 e 114 ocorrências (5%) o cluster 4.

Para o cenário de depois (2012-2016) foram registrados 2.235 registros de acidentes com VM. Verifica-se, neste cenário, que o cotovelo do gráfico ocorre para $k = 5$ e que para valores superiores a variação do WSS, Figura 4.5 (a), e seu respectivo valor médio, Figura 4.5 (b), não são significativos.

Figura 4.5 - Clustering para a base com VM para 2012-2016.

(a) WSS para diferentes valores de k com VM para 2012-2016.

(b) Média do WSS para diferentes valores de k com VM para 2012-2016.

(c) Valor médio da Silhueta para diferentes valores de k com VM para 2012-2016.

(d) Frequência das observações com $k=5$ para VM de 2012-2016.

Para $k = 5$ o valor médio da Silhueta foi de 0,86, Figura 4.5 (c), o que sugere que uma forte estrutura de grupo foi encontrada neste cenário. Na Figura 4.5 (d) verifica-se que do

total de 2235 registros de acidentes envolvendo VM no período de depois (2012 a 2016), 366 ocorrências (16%) compõem o cluster 1, 289 ocorrências (13%) o cluster 2, 794 ocorrências (36%) o cluster 3, 420 ocorrências (19%) o cluster 4 e 366 ocorrências o cluster 5 (16%).

4.2.2 Detecção de clusters de acidentes de trânsito com vítimas não fatais utilizando cenários antes e depois da intervenção

No cenário compreendido entre os anos de 2009 a 2012 foram detectados 616 registros de acidentes com VNF, que correspondem as ocorrências que resultaram em vítimas leves, moderadas e graves. Ao se analisar a variação do WSS, Figura 4.6 (a), e de seu valor médio, Figura 4.6 (b), verifica-se que a estrutura ótima de agrupamentos é observada para $k = 4$.

Figura 4.6 - Clustering para a base com VNF para 2009-2012.

(a) WSS para diferentes valores de k com VNF para 2009-2012.

(b) Média do WSS para diferentes valores de k com VNF para 2009-2012.

(c) Valor médio da Silhueta para diferentes valores de k com VNF para 2009-2012.

(d) Frequência das observações com $k = 4$ para VNF de 2009-2012.

Quando se analisa o valor médio da Silhueta para diferentes valores de k , verifica-se que para $k = 4$ o valor médio da Silhueta é de 0,94, Figura 4.6 (c), o que revela que os dados apresentam forte estrutura de grupo. Considerando os registros de VNF no trânsito,

Figura 52 (d), nota-se que o cluster 1 concentra 163 observações (26%), o cluster 2 concentra 124 observações (20%), o cluster 3 com 95 observações (15%) e o cluster 4 contém 234 observações (38%), Figura 4.6 (d).

Com os dados de 2012 a 2016, cenário de depois, foram observados 627 registros com VNF. Pela análise do WSS, Figura 4.7 (a), e de seu valor médio, Figura 4.7 (b), verifica-se que não há variação significativa na estrutura dos agrupamentos para $k > 4$. Portanto, o melhor valor para este parâmetro é $k = 4$.

Figura 4.7 - Clustering para a base com VNF para 2012-2016.

(a) WSS para diferentes valores de k com VNF para 2012-2016.

(b) Média do WSS para diferentes valores de k com VNF para 2012-2016.

(c) Valor médio da Silhueta para diferentes valores de k com VNF para 2012-2016.

(d) Frequência das observações com $k = 4$ para VNF de 2012-2016.

O valor médio da Silhueta para $k = 4$ é de 0,90, Figura 4.7 (c), o que indica que com este parâmetro tem-se uma estrutura robusta de clusters para os dados analisados. Pela Figura 4.7 (d), verifica-se que 108 ocorrências (17%) pertencem ao cluster 1, 191 ocorrências (30%) ao cluster 2, 204 ocorrências (33%) ao cluster 3 e 124 ocorrências (20%) ao cluster 4.

4.2.3 Detecção dos clusters de acidentes de trânsito com vítimas fatais utilizando cenários antes e depois da intervenção

A análise dos agrupamentos obtidos pelo método k – means com a base de acidentes de 2009 a 2012 considerando os registros com VF, indica que o algoritmo converge para $k = 3$. O WSS, Figura 4.8 (a), e seu respectivo valor médio, Figura 4.8 (b), não apresentam variações significativas para $k > 3$.

Figura 4.8 - Clustering para a base com VF para 2009-2012.

(a) WSS para diferentes valores de k com VF para 2009-2012.

(b) Média do WSS para diferentes valores de k com VF para 2009-2012.

(c) Valor médio da Silhueta para diferentes valores de k com VF para 2009-2012.

(d) Frequência das observações com $k = 3$ para VF de 2009-2012.

O valor médio da Silhueta para $k = 3$ é de 0,88, Figura 4.8 (c), o que indica que neste cenário os clusters apresentam forte estrutura de agrupamento. O cluster 1 apresenta 14 observações (37%), o cluster 2 apresenta 16 observações (42%) e o cluster 3 apresenta 8 observações (21%), conforme ilustra a Figura 4.8 (d).

Para o cenário de depois (2012 a 2016), o algoritmo converge com $k = 2$. A Figura 4.9 (a) e a Figura 4.9 (b) indicam que para $k > 2$ não há variações consideráveis, respectivamente, no WSS e em seu valor médio.

Figura 4.9 - Clustering para a base com VF para 2012-2016.

O valor médio da Silhueta para $k = 2$ é de 0,83 conforme ilustra a Figura 4.9 (c). A formação dos agrupamentos neste cenário indica que o cluster 1 concentra 10 observações (34%) e o cluster 2 apresenta 19 observações (66%), como se verifica na Figura 4.9 (d).

4.3 Redes complexas homogêneas aplicadas à detecção de comunidades de vítimas de acidentes no trânsito

As variáveis selecionadas na seção 4.1 foram analisadas visando reduzir a heterogeneidade dos dados, a partir da abordagem de construção de redes complexas homogêneas, pelo método dos k – vizinhos mais próximos, usando o algoritmo $k - NN$. Experimentos anteriores realizados com os bancos de dados explorados durante a pesquisa, sugerem que as medidas centralidade não são suficientemente eficientes no refinamento da detecção de possíveis agrupamentos (Chuerubim, Valejo e Silva, 2018). Desta forma, optou-se pela utilização da medida de desempenho Modularidade, na avaliação dos agrupamentos obtidos pela técnica de redes complexas homogêneas.

A nomenclatura $k - NN$ se refere as redes simétricas obtidas com base em similaridades. Os valores baixos de k geram redes desconexas e a partir do incremento

de k , verifica-se um aumento da densidade de arestas e, por conseguinte, o aumento da rede. As redes simétricas são as mais indicadas quando se deseja identificar de forma simultânea diferentes comunidades, uma vez que apresentam melhores propriedades de conectividade.

Neste trabalho foram geradas, por experimentação, redes com os valores de k iguais a 3, 5, 10, 15, 20, 25 e 30, para a validação do melhor valor a ser adotado como parâmetro k no algoritmo $k - NN$, sendo que para $k > 30$ não foram verificadas alterações significativas na rede. Para a detecção de comunidades nas bases de dados utilizadas, de acordo com o grau de severidade, foi utilizado o algoritmo *Fast Greedy* (Clauset *et al.*, 2004), que otimiza a medida de modularidade proposta por Newman (2004), utilizando como recurso computacional a busca “gulosa”, o que implica em um menor tempo de processamento. Embora existam métodos de detecção mais robustos e que sugerem um melhor particionamento dos dados em comunidades, os mesmos exigem um alto custo computacional, o que os tornam inviáveis para serem utilizados em grandes bases de dados como em Segurança Viária.

Desta forma, sendo um dos algoritmos mais populares e explorados na detecção de comunidades, optou-se em aplicar o *Fast Greedy*, uma vez que a detecção de comunidades contribuirá exclusivamente à análise exploratória e na redução da heterogeneidade da base de dados, com a indicação de relacionamentos ocultos ou subjetivos entre as variáveis que compõem as bases de dados analisadas. Os particionamentos obtidos com *Fast Greedy* e que apresentam os maiores valores de modularidade sugerem uma estrutura de rede mais robusta e, por consequência, em uma melhor divisão da rede em comunidades.

Os algoritmos de detecção de comunidades vêm apresentando melhores resultados quando comparados aos métodos tradicionais de agrupamentos como o $k - means$, pois enquanto o primeiro considera apenas a função de dissimilaridade aplicada por meio de uma função de distância, por exemplo, a distância euclidiana, o $k - NN$ analisa a vizinhança para cada valor de k selecionado. Embora os algoritmos de detecção de comunidades forneçam, em geral, a melhor detecção de comunidades, eles apresentam como desvantagem o tempo de processamento.

Com o propósito de comparar os agrupamentos gerados pelo $k - means$ em relação ao *Fast Greedy*, foram utilizadas as mesmas bases e variáveis da seção 4.2, e geradas comunidades com a mesma taxa de agrupamento no intervalo de 2 a 10, com a finalidade de detectar o melhor valor do parâmetro k para a construção da rede e definir o número

ótimo de agrupamentos, que aqui é denominado de comunidades (C). As comunidades serão discriminadas pela sigla C_i^n , em que i indica o índice da comunidade e n o número total de comunidades obtidas.

4.3.1 Detecção de comunidades de acidentes de trânsito com vítimas com danos materiais utilizando cenários antes e depois da intervenção

A análise de agrupamentos com a base de dados de 2009-2012 dos registros de acidentes com VM pela abordagem de redes complexas homogêneas foi realizada para diferentes valores de k , com a finalidade de verificar a estrutura mais robusta da rede utilizando o método $k-NN$ para diferentes valores de C , usando o algoritmo *Fast Greedy*, conforme ilustra a Figura 4.10.

Figura 4.10 - Construção das redes complexas para VM de 2009-2012 com a variação do parâmetro $k-NN$ e do número de comunidades (C).

Os resultados obtidos para a base analisada permitiram identificar que a estrutura mais robusta de rede é identificada para $k = 3$, com o maior valor médio da Modularidade de 0,76, Figura 4.11 (a). Portanto, o melhor parâmetro para a construção da rede é $3-NN$. A Figura 4.11 (b) apresenta a concentração do número de observações em cada comunidade obtida pelo algoritmo *Fast Greedy*. A $C_{i=1}^3$ agrupa 36% das observações (864 registros), a $C_{i=2}^3$ agrupa 18% das observações (429 registros) e a $C_{i=3}^3$ agrupa 47% das observações (1.140 registros).

Figura 4.11 - Comunidades em redes complexas considerando a base de VM para 2009-2012.

A rede $3-NN$ apresenta comunidades mais homogêneas para $C = 3$ e o aumento no particionamento dos dados com a geração de novas comunidades não resulta em variações significativas para $C > 3$. A Figura 4.12 (a) e a Figura 4.12 (b) indicam, respectivamente, a variação do WSS e de seu valor médio.

Figura 4.12 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VM de 2009-2012.

Para o cenário de depois (2012-216) com o banco de dados com VM verifica-se que o algoritmo $k-NN$ convergiu com $k = 25$ (Figura 4.13).

Figura 4.13 - Construção das redes complexas para VM de 2012-2016 com a variação do parâmetro $k-NN$ e do número de comunidades (C).

A rede $25-NN$ apresenta o valor médio da modularidade de 0,79, Figura 4.14 (a) e comunidades mais homogêneas para $C = 5$, sendo que para $C > 5$ não são observadas variações significantes. A $C_{i=1}^5$ agrupa 15% das observações (338 registros), a $C_{i=2}^5$ agrupa 15% das observações (341 registros), a $C_{i=3}^5$ agrupa 29% das observações (655 registros), a $C_{i=4}^5$ agrupa 25% das observações (550 registros) e a $C_{i=5}^5$ agrupa 16% das observações (351 registros), conforme ilustra a Figura 4.14 (b).

Figura 4.14 - Comunidades em redes complexas considerando a base de VM para 2012-2016.

(a) Valor médio da modularidade para diferentes valores do parâmetro $k-NN$ para VM de 2012-2016.

(b) Frequência das observações com $C = 5$ para VM de 2012-2016.

A Figura 4.15 (a) e a Figura 4.15 (b) indicam, respectivamente, a variação do WSS e de seu valor médio.

Figura 4.15 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VM de 2012-2016.

4.3.2 Detecção de comunidades de acidentes de trânsito com vítimas não fatais utilizando cenários antes e depois da intervenção

A análise da base de dados de 2009 a 2012 com os acidentes com VNF apresentou uma estrutura mais robusta de rede para $k=10$ (Figura 4.16).

Figura 4.16 - Construção das redes complexas para VNF de 2009-2012 com a variação do parâmetro $k-NN$ e do número de comunidades (C).

A rede $10-NN$ apresenta o valor médio da modularidade de 0,78, Figura 4.17 (a), e converge de forma ótima para $C=3$. A $C_{i=1}^3$ concentra 36% das observações (220 registros), a $C_{i=2}^3$ apresenta 28% das observações (170 registros) e a $C_{i=3}^3$ concentra 37% das observações (226 registros), conforme ilustra a Figura 4.17 (b).

Figura 4.17 - Comunidades em redes complexas considerando a base de VNF para 2009-2012.

Pode-se observar pela variação do WSS, Figura 4.18 (a), e do seu respectivo valor médio, Figura 4.18 (b), que para $C > 3$ a variação do WSS e de seu respectivo valor médio são insignificantes.

Figura 4.18 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VNF de 2009-2012.

A base de dados (2012-2016) para os acidentes com VNF apresenta a melhor estrutura de rede para $k=10$, conforme ilustra a Figura 4.19.

Figura 4.19 - Construção das redes complexas para VNF de 2012-2016 com a variação do parâmetro $k-NN$ e do número de comunidades (C).

A rede 10- NN apresenta Modularidade de 0,80, Figura 4.20 (a) e converge de forma ótima com $C=4$. A $C_{i=1}^4$ concentra 26% das observações (162 registros), a $C_{i=2}^4$ apresenta 30% das observações (191 registros), a $C_{i=3}^4$ concentra 22% das observações (140 registros) e a $C_{i=4}^4$ com 21% das observações (134 registros), conforme ilustra a Figura 4.20 (b).

Figura 4.20 - Comunidades em redes complexas considerando a base de VNF para 2012-2016.

(a) Valor médio da modularidade para diferentes valores do parâmetro $k-NN$ para VNF de 2012-2016.

(b) Frequência das observações com $C=4$ para VNF de 2012-2016.

O particionamento dos dados em grupos com $C>4$ não resulta em variações representativas, como se verifica pelo WSS e variação do valor médio, respectivamente, na Figura 4.21 (a) e na Figura 4.21 (b).

Figura 4.21 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VNF de 2012-2016.

(a) WSS para diferentes valores de C com VNF para 2012-2016.

(b) Média do WSS para diferentes valores de C com VNF para 2012-2016.

4.3.3 Detecção de comunidades de acidentes de trânsito com vítimas fatais utilizando cenários antes e depois da intervenção

A base de dados de 2009 a 2012 com os registros de VF apresentou a melhor estrutura de rede para $k = 3$, conforme ilustra a Figura 4.22.

Figura 4.22 - Construção das redes complexas para VF de 2009-2012 com a variação do parâmetro k -NN e do número de comunidades (C).

A rede 3-NN apresenta o valor médio da Modularidade de 0,57, Figura 4.23 (a) e converge de forma ótima para $C = 3$, Figura 4.23 (b). A $C_{i=1}^3$ concentra 39% das observações (15 registros), a $C_{i=2}^3$ apresenta 37% das observações (14 registros) e a $C_{i=3}^3$ concentra 24% das observações (9 registros).

Figura 4.23 - Comunidades em redes complexas considerando a base de VF para 2009-2012.

(a) Valor médio da modularidade para diferentes valores do parâmetro $k-NN$ para VF de 2009-2012.

(b) Frequência das observações com $C=3$ para VF de 2009-2012.

Em relação ao número de comunidades verifica-se que a variação do WSS, Figura 4.24 (a), e do seu valor médio, Figura 4.24 (b), são insignificantes a partir de $C = 3$, o que indica uma forte estrutura de grupos.

Figura 4.24 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VF de 2009-2012.

(a) WSS para diferentes valores de C com VF para 2009-2012.

(b) Média do WSS para diferentes valores de C com VF para 2009-2012.

A estrutura de redes com VF para o cenário de depois convergiu com $k = 5$ e, portanto, este é o melhor parâmetro para a definição dos agrupamentos, com Modularidade média de 0,57 (Figura 4.25).

Figura 4.25. Construção das redes complexas para VF de 2012-2016 com a variação do parâmetro $k-NN$ e do número de comunidades (C).

Para $C=4$ verifica-se que a $C_{i=1}^4$ concentra 4 observações (14%), a $C_{i=2}^4$ apresenta 9 observações (31%), a $C_{i=3}^4$ concentra 11 observações (38%) e a $C_{i=4}^4$ com 5 observações (17%), com valor médio da modularidade de 0,57, conforme ilustram respectivamente a Figura 4.26 (a) e a Figura 4.26 (b).

Figura 4.26 - Comunidades em redes complexas considerando a base de VF para 2012-2016.

(a) Valor médio da modularidade para diferentes valores do parâmetro $k-NN$ para VF de 2012-2016.

(b) Frequência das observações com $C=4$ para VF de 2012-2016.

Em termos de comunidade a estrutura ótima de agrupamentos é verificada quando $C=4$. Para valores superiores a $C=4$ não são observadas variações significantes, conforme se verifica na Figura 4.27 (a) e na Figura 4.27 (b).

Figura 4.27 - Variação do WSS considerando as comunidades obtidas em redes complexas para a base de VF de 2012-2016.

(a) WSS para diferentes valores de C com VF para 2012-2016.

(b) Média do WSS para diferentes valores de C com VF para 2012-2016.

4.4 Extração de regras de decisão a partir de agrupamentos de acidentes de vítimas no trânsito pela abordagem de CART

O Apêndice F apresenta a análise da estrutura topológica com PCA dos agrupamentos obtidos pelos métodos de clustering e redes complexas homogêneas.

O Apêndice G apresenta a representação da estrutura topológica dos agrupamentos resultantes dos métodos de particionamento k – means e *Fast Greedy*, com a finalidade de facilitar a visualização espacial dos agrupamentos obtidos com as bases de dados de registros de acidentes rodoviários com VM, VNF e VF para o cenário de intervenção antes e depois.

Para identificar os atributos mais discerníveis na formação e caracterização de cada agrupamento analisou-se a distribuição das variáveis em cada grupo, com a finalidade de identificar as características dominantes ou distorcidas, e com base nesta estratégia nomear os agrupamentos em função destes atributos. A partir dos agrupamentos detectados nas seções 4.2 e 4.3, foram extraídas regras de decisão usando a abordagem de estruturas em árvore, utilizando o algoritmo de classificação CART disponível no software SPSS, visando detectar padrões ou relacionamentos significativos entre as variáveis pertencentes a cada agrupamento.

Esta estratégia se deu em razão da redução da heterogeneidade dos dados por meio da aplicação preliminar dos algoritmos de agrupamento. Posteriormente, estes dados foram balanceados pela técnica SMOTE (Fawcett, 2016), para que o processo de classificação não sofresse a influência do desequilíbrio entre os valores das classes existentes nos bancos de dados, colaborando à obtenção de resultados mais consistentes, com boa acurácia e devidamente classificados.

A Tabela H.1 (Apêndice H) apresenta as regras de decisão obtidas a partir dos agrupamentos detectados pelos métodos de particionamento explorados nas seções 4.2 e 4.3. As árvores de decisão obtidas com o algoritmo de extração de regras de decisão CART bem como a respectiva matriz de confusão são apresentadas no Apêndice I.

Nota-se que, em geral, o método *Fast Greedy*, baseado na modelagem por redes complexas homogêneas, apresenta a maior taxa de acerto ou nível de confiança na extração de regras de decisão, bem como a maior acurácia geral no processo de classificação das observações atribuídas a cada agrupamento.

A base de dados com registro de VM para o cenário de antes e depois apresentam as melhores e maiores taxas de acerto com o algoritmo *Fast Greedy*, sendo de respectivamente 83% e 86%. Em termos de acurácia, para o cenário de antes, o resultado foi de 96%, e para o de depois, ele foi de 93%.

Dentre as características com maior probabilidade e as mais discerníveis entre os agrupamentos, apresentados no Apêndice J, os agrupamentos podem ser classificados genericamente com base nas características mais destoantes de cada agrupamento, conforme apresentado na Tabela 4.2.

Tabela 4.2: Discriminação geral dos agrupamentos com VM em função dos fatores destoantes e dominantes.

Banco	Agrup.	Classificação geral	Tamanho (%)
VM antes	1	Acidentes de trânsito do tipo colisão traseira e choque no sentido norte, causa provável motorista, perfil em nível, trechos com barreiras mistas, período da manhã ou da tarde	33
	2	Acidentes de trânsito do tipo colisão traseira e choque no sentido norte, causa provável motorista e outros, em locais onde não existe trevos, período da manhã ou tarde	33
	3	Acidentes de trânsito do tipo colisão traseira, choque, colisão lateral e engavetamento no sentido norte ou sul, causa provável motorista ou outros, perfil em nível ou aclone, período da manhã ou tarde	34
VM depois	1	Acidentes de trânsito do tipo colisão traseira, colisão lateral e engavetamento no sentido norte, condição de visibilidade parcial, perfil em nível e em locais onde não existem barreiras	21
	2	Acidentes de trânsito do tipo colisão traseira e choque no sentido norte ou sul, causa provável motorista, em locais com radar e perfil em aclone ou declive	21
	3	Acidentes de trânsito do tipo colisão traseira e choque no sentido norte ou sul, onde não há radar e trevos, trechos mistos com barreiras e perfil em aclone ou declive	20
	4	Acidentes de trânsito do tipo colisão traseira, choque e engavetamento no sentido norte ou sul, onde existe radar, causa provável motorista e outros, perfil em nível	21
	5	Acidentes de trânsito do tipo colisão traseira e choque no sentido norte, causa provável motorista ou outros, existe radar, perfil em nível e trechos mistos com barreiras	17

Por meio das regras de decisão aplicadas a cada agrupamento obtido com o *Fast Greedy* no cenário antes da intervenção com VM, usando a base de dados balanceada, identificou-se que os acidentes mais prováveis no agrupamento 1 são do tipo colisão traseira que ocorrem no sentido norte da via, com perfil em nível e no trecho compreendido entre o km 125 ao km 141,900. No agrupamento 2 e no agrupamento 3 os acidentes apresentam a maior probabilidade de ocorrência no sentido norte, em trechos com barreiras ou mistos, com a largura da faixa crítica superior a 2,96 m e velocidade máxima de 100 km/h.

Para o cenário de depois com VM com os agrupamentos obtidos com o *Fast Greedy* verifica-se que no agrupamento 1 os acidentes apresentam a maior probabilidade de ocorrência a partir do km 132,900 em trechos onde não existem barreiras e no sentido norte. No agrupamento 2 são mais prováveis os acidentes que ocorrem a partir do km 132,700 e no agrupamento 3 entre o km 125 ao km 132,700. No agrupamento 4 os acidentes são mais prováveis a partir do km 132,900, em trechos com barreiras ou mistos, no sentido norte da rodovia e associado a causa provável via/meio ambiente e outros.

Para a base de dados com registros de VNF, o cenário de antes da intervenção apresenta o melhor desempenho com o algoritmo *Fast Greedy*, com a taxa de acerto das regras de decisão de 100% e acurácia de 94%. Para o cenário de depois com VNF verifica-se que ambos os métodos de particionamento apresentam uma taxa de acerto de 100% e convergem com o mesmo número de agrupamentos ($k = 4$). No entanto, a acurácia da classificação foi superior para o método k – means com magnitude de 98%. A nomeação geral dos agrupamentos com base nas características mais destoantes apresentadas na Tabela J.2 (Apêndice J) é apresentada na Tabela 4.3.

Por meio das regras de decisão aplicadas a cada agrupamento obtido com o *Fast Greedy* no cenário antes da intervenção com VNF, usando a base de dados balanceada, permitiu identificar que os acidentes mais prováveis no agrupamento 1 apresentam a maior probabilidade de ocorrerem em trechos com barreiras ou mistos, no sentido Sul e com VDM de passeio ≤ 14.200 . No agrupamento 2 são mais prováveis os acidentes que ocorrem onde não existem barreiras e com a largura do acostamento superior a 2,67 m no agrupamento. E o agrupamento 3 os acidentes onde existe e não existe barreiras (mistos), sentido norte e largura da faixa crítica menor ou igual a 3,45 m.

Tabela 4.3: Discriminação geral dos agrupamentos com VNF em função dos fatores destoantes e dominantes.

Banco	Agrup.	Classificação geral	Tamanho (%)
VNF antes	1	Acidentes de trânsito do tipo colisão traseira, colisão lateral, choque e outros no sentido sul, causa provável motorista, com ou sem trevo, trechos com barreiras ou mistos e influenciados pela largura do canteiro central	33
	2	Acidentes de trânsito do tipo colisão traseira, colisão lateral e choque no sentido sul, condição de visibilidade boa, trechos sem trevo e influenciados pela largura do acostamento	33
	3	Acidentes de trânsito do tipo colisão traseira e choque no sentido norte, trechos sem trevo e com barreiras, sob influência da largura do acostamento	34
VNF depois	1	Acidentes de trânsito do tipo colisão traseira, choque, colisão lateral e outros no sentido Sul, causa provável via/meio ambiente e outros, condição de visibilidade boa, período noturno, velocidade máxima de 100 km/h e trechos com, sem ou barreiras mistas	25
	2	Acidentes de trânsito do tipo colisão traseira e choque no sentido norte ou sul, causa provável motorista ou outros, período da manhã ou noite, condição de visibilidade boa ou parcial, velocidade máxima de 100 km/h ou 110 km/h, trechos com barreiras mistas	25
	3	Acidentes de trânsito do tipo colisão traseira, choque, colisão lateral e outros no sentido norte ou sul, causa provável motorista, período da manhã ou noite, condição de visibilidade boa, veloc. 100 km/h ou 110 km/h, trechos com, sem ou barreiras mistas	25
	4	Acidentes de trânsito do tipo colisão traseira, choque e colisão lateral no sentido Norte, causa provável motorista ou outros, velocidade de 100 km/h, trechos com barreiras ou mistos	25

Para o cenário de depois com VNF, as regras de decisão obtidas a partir dos agrupamentos do *k-means* resultaram em acidentes mais prováveis no agrupamento 1 em segmentos da via superiores ao km 133, causa provável via/meio ambiente ou outros, no sentido sul da rodovia. O agrupamento 2 concentra os acidentes que ocorreram em trechos da via menores ou iguais ao 133 km. O agrupamento 3 concentra os acidentes que ocorreram em trechos da rodovia a partir do km 133, com causa provável motorista veículo. O agrupamento 4 concentra os acidentes que ocorreram a partir do km 133, porém exclusivamente no sentido norte da rodovia e associados a causa provável via/meio ambiente e outros.

Considerando a base de dados de acidentes com VF no cenário de antes verifica-se que o *Fast Greedy* apresentou a maior taxa de acerto na extração de regras (100%) e a melhor acurácia (83%). Para o cenário de depois com VF, o *k-means* resultou em uma taxa de acerto de 100% e acurácia de 86%.

A nomeação geral dos agrupamentos com base nas características mais destoantes, conforme apresentadas na Tabela J.3 (Apêndice J), é apresentada na Tabela 4.4.

Tabela 4.4: Discriminação geral dos agrupamentos com VF em função dos fatores destoantes e dominantes.

Banco	Agrup.	Classificação geral
VF antes	1	Acidentes do tipo atropelamento, no sentido norte ou sul, local do acidente alças de saída, onde existem trevos, condição de visibilidade parcial, perfil em aclave ou declive, período da manhã ou da noite
	2	Acidentes do tipo atropelamento, no sentido norte ou sul, causa provável outros, influenciados pela largura da faixa crítica, onde não existe trevo, trechos mistos com barreiras, período noturno
	3	Acidentes do tipo atropelamento e colisão frontal, causa provável outros, sentido norte, largura da faixa crítica, local faixa de rolamento, período da manhã, condição de visibilidade boa, parcial ou ruim, onde existe ou não iluminação
VF depois	1	Acidentes do tipo atropelamento, causa provável outros, em trechos com 2, 3 ou 4 faixas, período noturno, condição de visibilidade parcial, dispositivos de entrada/saída, velocidade máxima de 100 km/h a 110 km/h, perfil em nível, trechos mistos com barreiras
	2	Acidentes do tipo atropelamento e colisão traseira, pista com 3 faixas e velocidade máxima de 100 km/h

As regras de decisão obtidas com VF a partir do *Fast Greedy* para o período de antes revelam que o agrupamento 1 concentra os acidentes que ocorrem do km 125 ao km 138, e o agrupamento 3 os acidentes que ocorreram a partir do km 138. No período de depois, os agrupamentos obtidos com VF a partir do *k-means* sugerem que os acidentes mais prováveis do agrupamento 1 estão associados a causa provável veículo ou via/meio ambiente e que ocorreram do km 125 ao km 133,900. O agrupamento 2 e o agrupamento 3 estão associados a causa provável veículo ou via/meio ambiente e que ocorreram a partir do km 133,900 e o agrupamento 4 a causa provável motorista.

Para o cenário de depois, os agrupamentos obtidos com o *k-means* com dados de VF revelam que os acidentes do agrupamento 1 apresentam maior probabilidade de ocorrência do km 125 ao km 132,600. E os acidentes do agrupamento 2 apresentam maior probabilidade de ocorrência a partir do km 132,600.

4.5 Exclusão de variáveis por agregação ou eliminação em função do grau de relevância e do contexto da temática abordada

Por meio das regras de decisão exploradas na seção 4.4 verifica-se que as variáveis que mais contribuíram ao processo de extração das regras foram: tipo de acidente, condição da pista, quilometragem, sentido, acostamento, velocidade máxima, perfil, inclinação da pista, largura da faixa crítica, largura do acostamento, barreiras e VDMA de passeio.

No entanto, como abordado na seção 4.1.1 nem todas as variáveis do banco de dados apresentam alta comunalidade. Desta forma, com base nos experimentos realizados nas

seções anteriores, algumas variáveis serão eliminadas por não apresentarem variância significativa na formação dos agrupamentos, enquanto que outras serão agregadas em variáveis mais representativas e com maior comunalidade, com a finalidade de reduzir a dimensionalidade do banco de dados por seleção, agregação e exclusão das variáveis, conforme ilustra a Tabela 4.5.

Tabela 4.5: Redução da dimensionalidade do banco de dados por seleção, agregação e exclusão.

Nº	Variáveis	Mantidas	Excluída	Variável de agregação
1	Tipo de Acidente	X		
2	Causa provável		X	
3	Período do dia		X	
4	km	X		
5	Sentido			Lat e long
6	Local do acidente			Lat e long
7	Vítimas fatais			Acidente com vítimas
8	Vítimas com danos materiais			Acidente sem vítimas
9	Vítimas não fatais			Acidentes com vítimas
10	Condição de visibilidade	X		
11	Condição do tempo	X		
12	Condição da pista	X		
13	Traçado da pista	X		
14	Acostamento			Largura do acostamento
15	Largura do acostamento	X		
16	Luminosidade		X	
17	Radar	X		
18	Trevo			Lat e long
19	Ramo			Lat e long
20	Velocidade máxima	X		
21	Perfil da pista			Inclinação da pista
22	Inclinação da pista	X		
23	Sinalização de obras		X	
24	Sinalização Vertical			Lat e long
25	Sinalização horizontal			Lat e long
26	Latitude	X		
27	Longitude	X		
28	Número de faixas	X		
29	Largura da faixa crítica	X		
30	Canteiro central			Largura do canteiro central
31	Largura do canteiro central	X		
32	Barreiras	X		
33	VDM Passeio	X		
34	VDM Comercial	X		

Com base nestas considerações, o banco de dados foi reduzido a 20 variáveis, conforme ilustra a Tabela 4.6.

Tabela 4.6: Variáveis discriminatórias e essenciais ao desenvolvimento da abordagem proposta.

Nº	Variáveis Seleccionadas
1	Tipo de Acidente
2	km
3	Acidentes sem vítimas
4	Acidentes com vítimas
5	Condição de visibilidade
6	Condição do tempo
7	Condição da pista
8	Traçado da pista
9	Largura do acostamento
10	Radar
11	Velocidade máxima
12	Inclinação da pista
13	Latitude
14	Longitude
15	Número de faixas
16	Largura da faixa crítica
17	Largura do canteiro central
18	Barreiras
19	VDM Passeio
20	VDM Comercial

Conforme verificado na Tabela 4.6, em relação a severidade, os acidentes foram agrupados em acidentes sem vítimas (VM) e acidentes com vítimas (VNF e VF), com a finalidade de melhorar a estrutura dos dados em termos de comunidade e minimizar o problema de desbalanceamento com múltiplas classes. Em relação a largura do acostamento e a largura do canteiro central, quando forem iguais a zero significa que não há, respectivamente acostamento e canteiro central. Quanto a inclinação da pista até 2% será considerada em nível e acima deste valor quando positivo será acline e quando negativo declive. Além disso, será adotado o VDMA total que corresponde a soma dos veículos de passeio com os veículos comerciais.

4.6 Classificação de bases de severidade de acidentes rodoviários pela abordagem de redes neurais artificiais e redes neurais bayesianas

Com base nos experimentos realizados nas seções 4.2 e 4.3 verificou-se que o algoritmo de agrupamento que apresentou o melhor desempenho para as bases de dados de acidentes foi o algoritmo de redes complexas. Desta forma, a partir dos agrupamentos obtidos na seção 4.3 aplicou-se a classificação em redes utilizando a abordagem de ANN com MPL e BNN usando o classificador Naive Bayes.

O procedimento visou analisar a performance da classificação por ANNs e BNNs em dados de segurança viária. E partir da abordagem de classificação em redes que apresentou o melhor desempenho, extraiu-se regras de decisão mais prováveis que correlacionem um maior número de variáveis em relação a abordagem de CART (usada de forma exploratória na pesquisa), o que permite quantificar e qualificar a acidentalidade em diferentes segmentos da rodovia.

A Tabela 4.7 ilustra a precisão geral do processo de classificação considerando as bases de acidentes com VM, VNF e VF para o período de 2009 a 2012 e de 2012 a 2016. E a Tabela 4.8 o tempo de processamento usando cada abordagem.

Tabela 4.7: Performance dos classificadores ANN e BNN para bases de severidade de acidentes rodoviários.

2009-2012						2012-2016					
DM		VNF		VF		DM		VNF		VF	
ANN	BNN	ANN	BNN	ANN	BNN	ANN	BNN	ANN	BNN	ANN	BNN
100%	97,48%	100%	97,78%	100%	97,77%	99,82%	93,19%	99,86%	98,95%	100%	94,44%

Tabela 4.8: Tempo de processamento do processo de classificação usando ANN e BNN para bases de severidade de acidentes rodoviários.

2009-2012						2012-2016					
DM		VNF		VF		DM		VNF		VF	
ANN	BNN	ANN	BNN	ANN	BNN	ANN	BNN	ANN	BNN	ANN	BNN
2,97s	3,01s	0,6s	0,6s	0,04s	0,05s	20,19s	0,3s	3,85s	0,01s	0,18s	0,01s

Pode-se verificar por meio da Tabela 4.7 que o algoritmo de ANN com MPL tende a apresentar *overfitting* para as diferentes bases analisadas, o que pode ser verificado com a precisão geral de 100% para a maioria das bases de severidade analisadas, bem como pelas matrizes de confusão apresentadas nas Tabelas 4.9, 4.10, 4.11, 4.12, 4.13 e 4.14.

Em termos de processamento dos dados, verifica-se que para o período antes da intervenção ambas as abordagens são compatíveis. No entanto, para o período após a intervenção, em que o trecho ainda passa por obras, verifica-se que a BNN tem uma performance de processamento superior para os dados analisados.

Tabela 4.9: Matriz confusão obtida com ANN para VM no período de 2009 a 2012.

Comunidades	C_1	C_2	C_3
C_1	1140	0	0
C_2	0	1140	0
C_3	0	0	1140

Tabela 4.10: Matriz confusão obtida com BNN para VM no período de 2009 a 2012.

Comunidades	C_1	C_2	C_3
C_1	1098	0	42
C_2	0	1139	1
C_3	43	0	1097

Tabela 4.11: Matriz confusão obtida com ANN para VM no período de 2012 a 2016.

Comunidades	C_1	C_2	C_3	C_4	C_5
C_1	655	0	0	0	0
C_2	0	652	0	0	3
C_3	0	0	655	0	0
C_4	0	1	0	654	0
C_5	0	1	0	1	653

Tabela 4.12: Matriz confusão obtida com BNN para VM no período de 2012 a 2016.

Comunidades	C_1	C_2	C_3	C_4	C_5
C_1	654	0	0	1	0
C_2	0	634	0	5	16
C_3	0	0	655	0	0
C_4	0	148	0	493	14
C_5	0	13	0	26	616

Tabela 4.13: Matriz confusão obtida com ANN para VNF no período de 2009 a 2012.

Comunidades	C_1	C_2	C_3
C_1	226	0	0
C_2	0	226	0
C_3	0	0	226

Tabela 4.14: Matriz confusão obtida com BNN para VNF no período de 2009 a 2012.

Comunidades	C_1	C_2	C_3
C_1	220	4	2
C_2	9	217	0
C_3	0	0	226

Tabela 4.15: Matriz confusão obtida com ANN para VNF no período de 2012 a 2016.

Comunidades	C_1	C_2	C_3	C_4
C_1	191	0	0	0
C_2	0	191	0	0
C_3	0	0	191	0
C_4	1	0	0	190

Tabela 4.16: Matriz confusão obtida com BNN para VNF no período de 2012 a 2016.

Comunidades	C_1	C_2	C_3	C_4
C_1	191	0	0	0
C_2	0	191	0	0
C_3	0	0	189	2
C_4	4	0	2	185

Tabela 4.17: Matriz confusão obtida com ANN para VF no período de 2009 a 2012.

Comunidades	C_1	C_2	C_3
C_1	15	0	0
C_2	0	15	0
C_3	0	0	15

Tabela 4.18: Matriz confusão obtida com BNN para VF no período de 2009 a 2012.

Comunidades	C_1	C_2	C_3
C_1	15	0	0
C_2	0	14	1
C_3	0	0	15

Tabela 4.19: Matriz confusão obtida com ANN para VF no período de 2012 a 2016.

Comunidades	C_1	C_2	C_3	C_4
C_1	9	0	0	0
C_2	0	9	0	0
C_3	0	0	9	0
C_4	0	0	0	9

Tabela 4.20: Matriz confusão obtida com BNN para VF no período de 2012 a 2016.

Comunidades	C_1	C_2	C_3	C_4
C_1	9	0	0	0
C_2	0	9	0	0
C_3	0	0	9	0
C_4	0	2	0	7

As Tabelas 4.21 a 4.32 apresentam as estatísticas gerais de desempenho da classificação Verdadeiro Positivo (TP - *True Positive*), Verdadeiro Negativo (TN – *True Negative*), Falso Positivo (FP – *False Positive*) e Falso Negativo (FN – *False Negative*), considerando o erro da classificação obtida com o banco de dados original e com o banco de dados balanceado (Chen *et al.*, 2015). O TP corresponde ao número de positivos reais corretamente classificados. O TN define o número de reais negativos devidamente classificados. O FP expressa a porcentagem de instâncias realmente negativas e que foram

incorretamente classificadas como positivas, enquanto o FN indica o percentual de instâncias positivas classificadas erroneamente como negativas.

A taxa de variação de TP indica a proporção de instâncias previstas corretamente como positivas para todos os casos reais positivos, enquanto a taxa de variação de FP expressa a proporção de instâncias previstas e incorretamente classificadas como positivas para todos os casos reais negativos (Chen *et al.*, 2015).

A precisão corresponde a proporção de instâncias devidamente classificadas entre todas as instâncias e fornece informações gerais sobre o desempenho do classificador. No entanto, esta métrica não é eficiente na avaliação de banco de dados desequilibrados (Mujalli, López & Garach, 2016).

Tabela 4.21: Exatidão do classificador ANN na modelagem da severidade da base de VM no período de 2009 a 2012.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	1,00	0,00	1,00	1,00	1,00
C_2	1,00	0,00	1,00	1,00	1,00
C_3	1,00	0,00	1,00	1,00	1,00

Tabela 4.23: Exatidão do classificador ANN na modelagem da severidade da base de VM no período de 2012 a 2016.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	1,00	0,00	1,00	1,00	1,00
C_2	0,99	0,01	1,00	1,00	1,00
C_3	1,00	0,00	1,00	1,00	1,00
C_4	0,99	0,00	1,00	1,00	1,00
C_5	1,00	0,00	0,99	1,00	1,00

Tabela 4.25: Exatidão do classificador ANN na modelagem da severidade da base de VNF no período de 2009 a 2012.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	1,00	0,00	1,00	1,00	1,00
C_2	1,00	0,00	1,00	1,00	1,00
C_3	1,00	0,00	1,00	1,00	1,00

Tabela 4.27: Exatidão do classificador ANN na modelagem da severidade da base de VNF no período de 2012 a 2016.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	1,00	0,00	0,99	1,00	1,00
C_2	1,00	0,00	1,00	1,00	1,00
C_3	1,00	0,00	1,00	1,00	1,00
C_4	0,99	0,00	1,00	1,00	1,00

Tabela 4.22: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VM no período de 2009 a 2012.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	0,96	0,02	0,96	0,96	0,99
C_2	0,99	0,00	1,00	1,00	1,00
C_3	0,96	0,02	0,96	0,96	0,99

Tabela 4.24: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VM no período de 2012 a 2016.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	1,00	0,00	1,00	0,99	1,00
C_2	0,97	0,06	0,80	0,87	0,99
C_3	1,00	0,00	1,00	1,00	1,00
C_4	0,75	0,01	0,94	0,84	0,98
C_5	0,94	0,01	0,95	0,95	0,99

Tabela 4.26: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VNF no período de 2009 a 2012.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	0,97	0,02	0,96	0,97	1,00
C_2	0,96	0,01	0,98	0,97	1,00
C_3	1,00	0,04	0,99	1,00	1,00

Tabela 4.28: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VNF no período de 2012 a 2016.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	1,00	0,00	0,98	0,99	1,00
C_2	1,00	0,00	1,00	1,00	1,00
C_3	1,00	0,00	1,00	0,99	0,99
C_4	0,97	0,00	0,99	0,98	0,99

Tabela 4.29: Exatidão do classificador ANN na modelagem da severidade da base de VF no período de 2009 a 2012.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	1,00	0,00	1,00	1,00	1,00
C_2	1,00	0,00	1,00	1,00	1,00
C_3	1,00	0,00	1,00	1,00	1,00

Tabela 4.31: Exatidão do classificador ANN na modelagem da severidade da base de VF no período de 2012 a 2016.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	1,00	0,00	1,00	1,00	1,00
C_2	1,00	0,00	1,00	1,00	1,00
C_3	1,00	0,00	1,00	1,00	1,00
C_4	1,00	0,00	1,00	1,00	1,00

Tabela 4.30: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VF no período de 2009 a 2012.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	1,00	0,00	1,00	1,00	1,00
C_2	0,93	0,00	1,00	0,97	1,00
C_3	1,00	0,03	0,94	0,97	1,00

Tabela 4.32: Exatidão do classificador Naive Bayes na modelagem da severidade da base de VF no período de 2012 a 2016.

Comunidades	TP	FP	Precisão	Medida-F	AUC
C_1	1,00	0,00	1,00	1,00	1,00
C_2	1,00	0,07	0,82	0,90	0,96
C_3	1,00	0,00	1,00	1,00	1,00
C_4	0,78	0,00	1,00	0,87	0,93

Verifica-se que para os bancos de severidade analisados as abordagens estatisticamente foram compatíveis. No entanto, a classificação por ANN foi mais otimista, acertando, em geral, 100% das instâncias analisadas.

Todavia, como abordado na seção 2.3.4 as BNNs consideram os vieses existentes entre as camadas utilizadas na transferência dos dados, bem como o peso de cada parâmetro, sendo mais adequada ao mapeamento entre causas e efeitos. O peso inicial dos parâmetros na abordagem por BNNs é definido a partir das distribuições amostrais construídas para as variáveis de interesse e na probabilidade “a priori” (Lawrence *et al.*, 1998; Permai, Ulama e Iriawan, 2015).

Nas BNNs o conhecimento “a priori” de todas variáveis que compõem a base de dados é utilizado em conjunto na arquitetura em redes (relações de independência condicionada), por meio de uma função de probabilidade, o que possibilita uma aproximação eficiente do classificador evitando ou minimizando o *overfitting*, que é um problema recorrente em métodos de aprendizado supervisionado por ANNs.

Nas ANNs, por sua vez, o procedimento de determinação da complexidade do modelo é mais custoso, pois implica em uma maior dificuldade em se determinar o número ideal de parâmetros necessários a modelagem, bem como na seleção de ferramentas para análise dos resultados. Desta forma, na prática deve-se adotar aproximações bayesianas para gerenciar tal limitação, por meio da definição a priori dos parâmetros de maior peso estatístico (hiperparâmetros).

A complexidade do modelo resultante será igual a média de todas as complexidades definidas pelas probabilidades “a posteriori”, determinadas a partir da amostragem dos dados de entrada.

A aproximação bayesiana permite realizar inferências em ANNs, mesmo quando alguma informação “a priori” for inconsistente ou inexistente. A análise das amostras possibilita a obtenção da distribuição dos parâmetros “a posteriori”, refinando-se, assim, o conhecimento “a priori”, procedimento deste denominado de “aprendizado bayesiano”.

4.6.1 Regras de classificação obtidas pela abordagem de redes bayesianas a partir das bases de severidade de acidentes rodoviários

A partir da classificação em redes usando a abordagem por BNNs foram extraídas regras de classificação para as diferentes bases de severidades de dados estruturados em redes. Desta forma, foram selecionadas as regras de classificação de maior probabilidade de ocorrência, excluindo os FPs, conforme verificado nas matrizes de confusão apresentadas na seção 4.6.

As observações mais prováveis de ocorrência dos acidentes foram analisadas de forma microscópica considerando as características dominantes em cada comunidade de acidentes.

Acidentes com vítimas materiais no período de antes

A Figura 4.28 apresenta a variação da AUC para as observações mais prováveis da C_1 considerando a base de severidade com VM para o período de “antes” (2009-2012).

Figura 4.28 - Variação da AUC para a C_1 considerando a base de severidade com VM para o período de “antes” (2009-2012).

A contribuição das variáveis na ocorrência dos acidentes pertencentes a C_1 considerando a base de severidade com VM para o período de “antes” (2009-2012) é apresentada na Figura K.1 do Apêndice K (em CD-ROM). Os acidentes foram observados com maior frequência nos marcos quilométricos km 129, km 131, km 135, km 139 e km

145 no sentido norte, Figura K.1 (a). Nestes marcos foram verificados os maiores VMDs, Figura K.1 (b).

Os tipos de acidentes mais frequentes na C_1 foram colisão traseira, engavetamento e choque, Figura K.1 (c). Estes acidentes ocorreram predominantemente em condição de tempo bom, condição da pista seca e condição de visibilidade boa ou parcial, Figura K.1 (d). Em relação ao perfil da rodovia estes acidentes foram observados predominantemente em trechos em nível, com inclinação entre -2% a 2%, conforme ilustra a Figura K.1 (e). Foram observados com maior frequência em segmentos da rodovia em que a largura média do canteiro central é de 11,72 m, Figura K.1 (f), a largura média do acostamento de 3 m, Figura K.1 (g), em trechos com 3 faixas, com largura média da faixa crítica de 3 m e velocidade máxima de 100 km/h, sendo mais frequentes em trechos com traçado reto, Figura K.1 (h).

Na segunda comunidade C_2 os fatores chaves foram selecionados com uma AUC média de 0,99, conforme ilustra a Figura 4.29.

Figura 4.29 - Variação da AUC para a C_2 considerando a base de severidade com VM para o período de “antes” (2009-2012).

A contribuição das variáveis na ocorrência dos acidentes pertencentes a C_2 , considerando a base de severidade com VM para o período de “antes” (2009-2012), é apresentada na Figura K.2 do Apêndice K (em CD-ROM). Esta comunidade representa os acidentes que ocorreram predominantemente no sentido norte e com maior frequência nos marcos quilométricos km 126, km 130, km 132, km 133, km 139 e km 141, conforme ilustra a Figura K.2 (a). Nestes marcos são verificados os maiores VDMs de veículos, Figura K.2 (b). Os tipos de acidentes mais frequentes foram colisão traseira, colisão lateral e choque, em trechos mistos com barreiras, Figura K.2 (c).

Estes acidentes ocorreram predominantemente em condição de tempo bom, condição de pista seca e condição de visibilidade boa ou parcial, Figura K.2 (d). Quanto ao perfil estes acidentes ocorreram predominantemente em trechos em nível, com inclinação de variando de -2% a 2%, Figura K.2 (e). Os acidentes ocorreram em trechos em que a largura do canteiro central é de 13 m, Figura K.2 (f), com largura média do acostamento de 3 m, Figura K.2 (g), em trechos com 3 faixas e com a largura média da faixa crítica de 3 m, ocorreram predominantemente em trechos com traçado reto, Figura K.2 (h).

A Figura 4.30 apresenta a variação média de AUC 0,99 para as observações da comunidade C_3 .

Figura 4.30 - Variação da AUC para a C_3 considerando a base de severidade com VM para o período de “antes” (2009-2012).

A contribuição das variáveis na ocorrência dos acidentes pertencentes a C_3 , considerando a base de severidade com VM para o período de “antes” (2009-2012), é apresentada na Figura K.3 do Apêndice K (em CD-ROM). Nesta comunidade os acidentes mais frequentes foram observados no sentido norte no km 128, km 129, km 134, km 144 e km 145, enquanto no sentido sul no km 126, km 128, km 133 e km 145, Figura K.3 (a). Nestes marcos foram verificados os maiores VDMs de veículos, Figura K.3 (b). Os tipos de acidentes mais frequentes foram colisão traseira, engavetamento e choque, Figura K.3 (c).

Em ambos os sentidos da rodovia os acidentes ocorreram com maior frequência em condição de tempo bom, condição de pista seca ou molhada, condição de visibilidade boa ou parcial, Figura K.3 (d). No sentido norte há uma predominância de acidentes que ocorreram em nível, com inclinação da pista entre -2% a 2%, ao passo que no sentido sul, os acidentes ocorreram predominantemente em trechos com aclive (com inclinação média de 3,32% ou declive (com inclinação média de 4%), Figura K.3 (e).

Os acidentes no trecho norte foram observados com maior frequência em segmentos da rodovia com a largura média do canteiro central de 13 m, e no sentido sul com valor médio de 12 m, Figura K.3 (f). No sentido norte ocorreram em trechos com a largura média do acostamento de 3 m e no sentido sul de 3,82 m, Figura K.3 (g). Estes acidentes ocorreram com maior frequência no sentido norte em trechos com 3 faixas, com a largura média faixa crítica de aproximadamente 3 m, com velocidade máxima de 80 km/h, 90 km/h, 100 km/h ou 110 km/h. No sentido sul em segmentos da rodovia com 2 ou 3 faixas, com limite de velocidade máxima de 100 km/h ou 110 km/h, com o valor médio da largura da faixa crítica de, aproximadamente, de 3 m, em trechos predominantemente com traçado reto em ambos os sentidos da rodovia, Figura K.3 (h).

Acidentes com vítimas materiais no período de depois

A Figura 4.31 apresenta a variação média de AUC 0,99 para as observações da comunidade C_1 .

Figura 4.31 - Variação da AUC para a C_1 considerando a base de severidade com VM para o período de “depois” (2012-2016).

Na comunidade C_1 considerando a base de VM para o período de “depois, Figura K.4 do Apêndice K (em CD-ROM), verifica-se que os acidentes ocorreram no sentido norte, sendo mais frequentes no km 143 e no km 145, Figura K.4 (a). Nestes marcos quilométricos são verificados os maiores VDMs de veículos para o período analisado, Figura K.4 (b). Os tipos de acidentes mais frequentes foram colisão traseira, engavetamento, choque e colisão lateral, Figura K.4 (c). Os mesmos ocorreram em condição de tempo bom ou chuva, condição de pista seca ou molhada, e condição de visibilidade boa ou parcial, Figura K.4 (d).

Os acidentes da C_1 com VM para o período de “depois” ocorreram predominantemente no sentido norte, em trechos em nível, com inclinação entre -2% a 2%, e em acive, com inclinação entre 2,93% a 4,53%, Figura K.4 (e). Nestes trechos o valor médio da largura do canteiro central é de 13 m, Figura K.4 (f), a largura do acostamento de 3 m, Figura K.4 (g). Foram observados com maior frequência em trechos com 2 faixas, limite de velocidade máxima de 100km/h e largura média da faixa crítica de 3 m, em trechos exclusivamente retos, Figura K.4 (h).

A Figura 4.32 apresenta a variação média de AUC 0,99 para as observações da comunidade C_2 .

Figura 4.32 - Variação da AUC para a C_2 considerando a base de severidade com VM para o período de “depois” (2012-2016).

A contribuição de cada variável na ocorrência dos acidentes com VM na comunidade C_2 no período de “depois” é apresentada na Figura K.5 do Apêndice K (em CD-ROM). Nesta comunidade a maior concentração de acidentes ocorreu no sentido norte nos marcos quilométricos km 134 e km 137, enquanto no sentido sul no km 135, km 136, km 139 e km 141, Figura K.5 (a). Nestes marcos quilométricos foram verificados os maiores VDMs de veículos, conforme ilustra a Figura K.5 (b). Os tipos de acidentes mais frequentes no sentido norte foram choque, engavetamento e colisão traseira, enquanto no sentido sul os acidentes do tipo choque e colisão traseira, Figura K.5 (c).

Em ambos os sentidos da rodovia os acidentes foram observados predominantemente em condição de tempo bom, condição de pista seca e condição de visibilidade boa ou parcial, Figura K.5 (d). No sentido norte os acidentes ocorreram predominantemente em nível (-2% a 2%) e no sentido sul em declive (com inclinação média de 3%), Figura K.5 (e). Estes acidentes ocorreram em trechos em que a largura média do canteiro central é

de 13 m, Figura K.5 (f). A largura do acostamento observada em ambos os sentidos da rodovia foi de 3 m, Figura K.5 (g).

Os acidentes observados na comunidade C_2 no período de “depois” ocorreram predominantemente em segmentos da rodovia com 2 faixas, com a largura média da faixa crítica de 3 m e limite de velocidade de 90 km/h, com traçado reto predominantemente em ambos os sentidos da rodovia, conforme ilustra a Figura K.5 (h).

A Figura 4.33 apresenta a variação média de AUC 0,99 para as observações da comunidade C_3 .

Figura 4.33 - Variação da AUC para a C_3 considerando a base de severidade com VM para o período de “depois” (2012-2016).

Na comunidade C_3 , Figura K.6 do Apêndice K (em CD-ROM), os acidentes mais frequentes no sentido norte foram observados km 126, km 129 e km 132, enquanto que no sentido sul foram observados nos marcos quilométricos km 126, km 127, km 128, km 131, km 132 e km 133, Figura K.6 (a). Nestes marcos foram verificados os maiores VDMs de veículos, Figura K.6 (b). Os tipos de acidentes mais frequentes em ambos os sentidos foram choque, colisão traseira, colisão lateral e engavetamento, Figura K.6 (c). Estes acidentes foram observados em condição de tempo bom ou com chuva, condição de pista seca ou molhada e condição de visibilidade boa ou parcial, Figura K.6 (d).

Em ambos os sentidos da rodovia os acidentes ocorreram em trechos em nível (inclinação entre -2% a 2%), em aclave (inclinação maior do que 2%) ou declive (inclinação menor que -2%), Figura K.6 (e). Estes acidentes ocorreram tanto no sentido norte ou sul em trechos da rodovia em que a largura média do canteiro do canteiro central era de 11 m, Figura K.6 (f), e a largura média do acostamento de 3 m, Figura K.6 (g).

No sentido norte ou sul ocorreram exclusivamente em trechos com 2 faixas. No sentido norte foram observados com maior frequência em trechos com limite de velocidade de 110 km/h, com a largura da faixa crítica maior do que 3 m. No sentido sul, ocorreram predominantemente em trechos com limite de velocidade de 90 km/h e com a largura da

faixa crítica menor ou igual a 3 m. Os acidentes ocorreram predominantemente em trechos retos da rodovia, Figura K.6 (h).

Na comunidade C_4 , ilustrada na Figura 4.34, verifica-se que a variação média de AUC é de 0,99 para o conjunto de observações.

Figura 4.34 - Variação da AUC para a C_4 considerando a base de severidade com VM para o período de “depois” (2012-2016).

Na comunidade C_4 com VM no período de “depois” se concentraram no sentido norte nos marcos quilométricos km 134, km 135, km 141, km 142, km 144 e km 145, enquanto no sentido sul nos marcos quilométricos km 135, km 137 ao km 139 e no km 141, Figura K.7 (a). Nestes marcos foram verificados os maiores VDMs de veículos, Figura K.7 (b).

Em ambos os sentidos da rodovia os tipos de acidentes mais frequentes foram colisão traseira, colisão lateral, choque e engavetamento, Figura K.7 (c). Tanto no sentido norte quanto no sentido sul ocorreram predominantemente em condição de tempo bom ou com chuva, condição de pista seca ou molhada e condição de visibilidade boa ou parcial, Figura K.7 (d).

No trecho norte os acidentes ocorreram predominantemente em nível (com inclinação entre -2% a 2%), enquanto no sentido sul em trechos com auge (inclinação maior que 2%) ou declive (inclinação menor que -2%), Figura K.7 (e). Em ambos os sentidos da rodovia os acidentes ocorreram predominantemente em trechos em que a largura do canteiro central era de 11 m, Figura K.7 (f), e a largura do acostamento de 3 m, Figura K.7 (g).

No sentido norte os acidentes ocorreram predominantemente em trechos com 2 faixas, com limite de velocidade de 80 km/h ou 90 km/h, e com a largura média da faixa crítica de 3 m. Enquanto que no sentido sul, os acidentes também ocorreram em trechos com 2 faixas, com limite de velocidade de 90 km/h e largura média da faixa crítica de 3 m. Em

ambos os sentidos da rodovia correram predominantemente em trechos retos, Figura K.7 (h).

Na comunidade C_5 , ilustrada na Figura 4.35, verifica-se que a variação média de AUC é de 0,99 para o conjunto de observações.

Figura 4.35 - Variação da AUC para a C_5 considerando a base de severidade com VM para o período de “depois” (2012-2016).

Na comunidade C_5 , Figura K.8 do Apêndice K (em CD-ROM), os acidentes ocorreram exclusivamente no sentido norte e com maior frequência nos marcos quilométricos km 135, km 137, km 139, km 142 e km 143, Figura K.8 (a). Nestes marcos são verificados os maiores VDMs de veículos para o período observado, Figura K.8 (b).

Os tipos de acidentes mais frequentes neste trecho são do tipo choque, colisão traseira, engavetamento e outros, que ocorreram predominantemente em trechos mistos com barreiras, Figura K.8 (c). Estes acidentes ocorreram predominantemente em condição de tempo bom ou com chuva, condição da pista seca ou molhada e condição de visibilidade boa ou parcial, Figura K.8 (d).

Os acidentes ocorreram em trechos em nível, com inclinação entre -2% a 2%, Figura K.8 (e). Os acidentes ocorreram predominantemente em trechos da rodovia com a largura do canteiro central de 13 m, Figura K.8 (f), e largura do acostamento de 3 m, Figura K.8 (g). Foram observados em trechos da rodovia com 2 faixas, com limite de velocidade de 90 km/h, e com a largura média da faixa crítica de 3 m. No sentido norte ocorreram em trechos retos ou curvos, enquanto no sentido sul em trechos retos, Figura K.8 (h).

Acidentes com vítimas não fatais no período de antes

Na comunidade C_1 , ilustrada na Figura 4.36, verifica-se que a variação média de AUC é de 0,99 para o conjunto de observações.

Figura 4.36 - Variação da AUC para a C_1 considerando a base de severidade com VNF para o período de “antes” (2009-2012).

No sentido norte os acidentes da C_1 com VNF no período de “antes”, Figura K.9 do Apêndice K (em CD-ROM), ocorreram predominantemente no km 128 e no km 145, enquanto que no sentido sul nos marcos quilométricos km 128, km 131 ao km 134, km 137 ao km 139, km 141, km 142 e km 145, Figura K.9 (a). Nestes marcos foram verificados os maiores VDMs de veículos, Figura K.9 (b).

Em ambos os sentidos da rodovia os acidentes na C_1 com VNF no período de “antes” ocorreram em trechos com barreiras ou mistos, sendo os tipos de acidentes mais frequentes do tipo choque, colisão lateral, colisão traseira, tombamento e outros, Figura K.9 (c). Os acidentes do tipo atropelamento envolvendo VNF no período de “antes” na C_1 foram mais frequentes no sentido sul no km 142 da rodovia, conforme ilustra a Figura K.9 (c).

Os acidentes na C_1 com VNF no período de “antes” ocorreram predominantemente em condição de tempo bom, com chuva ou garoa, condição de pista seca ou molhada, e condição de visibilidade boa ou parcial, Figura K.9 (d). Em ambos os sentidos da rodovia foram observados em trechos em nível (inclinação entre -2% a 2%), em alicive (inclinação maior que 2%) ou declive (inclinação inferior a -2%), Figura K.9 (e). Ocorreram tanto no sentido norte ou sul em trechos com a largura média do canteiro central de 13 m, Figura K.9 (f), e a largura média do acostamento de 3 m, Figura K.9 (g).

Em ambos os sentidos da rodovia ocorreram predominantemente em trechos com 2 ou 3 faixas. Sendo mais frequentes no sentido norte em segmentos com limite de velocidade de 80 km/h (largura média da faixa crítica de 3 m) ou 100 km/h (com largura média da faixa crítica de 3,35 m). No sentido sul, foram observados com maior frequência em trechos com limite de velocidade de 90 km/h (largura média da faixa crítica de 3 m) ou 110 km/h (largura média da faixa crítica de 3,35 m). No sentido norte os acidentes da C_1

com VNF no período de “antes” ocorreram em trechos retos ou curvos, enquanto no sentido sul exclusivamente em trechos retos. No sentido norte foram observados em trechos retos, enquanto no sentido sul em trechos retos ou curvos, Figura K.9 (h).

O conjunto de observações da comunidade C_2 com VNF no período de “antes”, Figura K.10 do Apêndice K (em CD-ROM), apresentaram todas $AUC = 1$. Nesta comunidade os acidentes foram observados com maior frequência no sentido norte no km 144, enquanto no sentido sul nos marcos quilométricos km 130, km 135, km 141 ao km 144, Figura K.10 (a). Nestes marcos foram verificados os maiores VDMs de veículos, Figura K.10 (b). Em ambos os sentidos da rodovia os acidentes ocorreram predominantemente em trechos sem barreiras. Os tipos de acidentes mais frequentes foram colisão traseira, colisão lateral, choque e outros, Figura K.10 (c).

Os acidentes na comunidade C_2 ocorreram em condição de tempo bom ou com chuva, condição de pista seca ou molhada, e condição de visibilidade boa ou parcial, Figura K.10 (d). Estes acidentes ocorreram predominantemente em nível, com inclinação da pista entre -2% a 2%, Figura K.10 (e). No sentido norte os acidentes ocorreram predominantemente em trechos da rodovia com a largura média do canteiro central de 13 m, ao passo que no sentido sul com a largura média do canteiro central de 11 m, Figura K.10 (f).

Em ambos os sentidos da rodovia os acidentes ocorreram em trechos em que a largura média do acostamento era de 3 m, Figura K.10 (g). No sentido norte os acidentes foram observados exclusivamente em trechos da rodovia com 3 faixas, com limite de velocidade de 100 km/h e com a largura média da faixa crítica de 3,60 m. Já no sentido sul, os acidentes foram observados em trechos com 2 ou 3 faixas, com limite de velocidade de 100 km/h ou 110 km/h, e com a largura média da faixa crítica de 3 m. Em ambos os sentidos da rodovia ocorreram predominantemente em trechos retos, Figura K.10 (h).

Na comunidade C_3 , ilustrada na Figura 4.37, verifica-se que a variação média de AUC é de 0,99 para o conjunto de observações.

Figura 4.37 - Variação da AUC para a C_3 considerando a base de severidade com VNF para o período de “antes” (2009-2012).

Na comunidade C_3 com VNF “antes”, Figura K.11 do Apêndice K (em CD-ROM), os acidentes ocorreram exclusivamente no sentido norte da rodovia, nos marcos quilométricos km 126, km 129, km 131 ao km 135, km 137, km 139, km 141 e km 142, Figura K.11 (a). Nestes marcos quilométricos foram verificados os maiores VDMs de veículos para o período observado, Figura K.11 (b).

Estes acidentes ocorreram em trechos sem barreiras, com barreiras e mistos. Os tipos de acidentes mais frequentes foram choque, colisão traseira, engavetamento e outros, Figura K.11 (c). Foram observados predominantemente em condição de tempo bom ou com chuva, condição de pista seca ou molhada, e condição de visibilidade boa ou parcial, Figura K.11 (d).

Os acidentes da C_3 com VNF “antes” ocorreram em trechos em nível (inclinação variando entre -2% a 2%), aclive (inclinação maior que 2%) e declive (menor que -2%), Figura K.11 (e). Ocorreram em trechos em que a largura do canteiro central era de 11 m, Figura K.11 (f), e largura do acostamento média de 3 m, Figura K.11 (g). Ocorreram em trechos com 3 faixas, com limite de velocidade de 80 km/h a 110 km/h, com a largura média da faixa crítica de 3 m. No sentido norte ocorreram exclusivamente em trechos curvos e no sentido sul em trechos retos, Figura K.11 (h).

Acidentes com vítimas não fatais no período de depois

Na comunidade C_1 , ilustrada na Figura 4.38, verifica-se que a variação média de AUC é de 0,99 para o conjunto de observações.

Figura 4.38 - Variação da AUC para a C_1 considerando a base de severidade com VNF para o período de “depois” (2012-2016).

Na comunidade C_1 com VNF para o período de “depois”, Figura K.12 do Apêndice K (em CD-ROM), no trecho norte, os acidentes ocorreram com maior frequência do km 141 ao km 145, enquanto que no sentido sul os acidentes foram observados com maior frequência do km 143 ao km 145, Figura K.12 (a). Nestes trechos foram observados os maiores VDMs de veículos, Figura K.12 (b). Os tipos de acidentes mais frequentes em ambos os sentidos foram colisão traseira, colisão lateral e choque, em trechos com ou sem barreiras, Figura K.12 (c).

Em ambos os sentidos da rodovia os acidentes ocorreram em condição de tempo bom ou com chuva, condição de pista seca ou molhada, e condição de visibilidade boa ou parcial, Figura K.12 (d). No sentido norte estes acidentes ocorreram predominantemente em nível (inclinação entre -2% a 2%), ao passo que no sentido sul em trechos com aclive (inclinação maior que 2%), Figura K.12 (e).

Tanto no sentido norte quanto no sentido sul, os acidentes foram observados em trechos da rodovia com a largura média do canteiro central de 11 m, Figura K.12 (f), e largura do acostamento de 3 m, Figura K.12 (g). Em ambos os sentidos os acidentes ocorreram exclusivamente em trechos com 2 faixas, com limite de velocidade de 80 km/h a 90 km/h, com a largura média da faixa crítica de 3 m. Em ambos os sentidos da rodovia ocorreram exclusivamente em trechos retos, Figura K.12 (h).

Na comunidade C_2 , ilustrada na Figura 4.39, verifica-se que a variação média de AUC é de 0,99 para o conjunto de observações.

Figura 4.39 - Variação da AUC para a C_2 considerando a base de severidade com VNF para o período de “depois” (2012-2016).

Na comunidade C_2 considerando VNF no período de “depois”, Figura K.13 do Apêndice K (em CD-ROM), os acidentes foram observados com maior frequência no sentido norte no km 134, km 135, km 137 e km 138, enquanto que no sentido sul no km 134, km 135, km, 139, km 141 e km 142, Figura K.13 (a). Nestes marcos foram verificados os maiores VDMs de veículos, Figura K.13 (b).

No sentido norte os acidentes ocorreram em trechos com barreiras ou mistos, enquanto no sentido sul predominantemente em trechos mistos com barreiras. Os tipos de acidentes mais frequentes foram colisão traseira, choque e colisão lateral, Figura K.13 (c). Os acidentes foram observados em ambos os sentidos da rodovia em condição de tempo bom, condição de pista seca e condição de visibilidade boa, Figura K.13 (d).

Tanto no sentido norte ou sul os acidentes ocorreram predominantemente em trechos, em nível (inclinação entre -2% a 2%), aclave (inclinação superior a 2%) e declive (inclinação inferior a -2%), Figura K.13 (e). Em ambos os sentidos da rodovia os acidentes ocorreram predominantemente com a largura do canteiro central de 11 m, Figura K.13 (f), e a largura do acostamento de 3 m, Figura K.13 (g).

Os acidentes ocorreram exclusivamente em trechos com 3 faixas, com limite de velocidade de 90 km/h a 110 km/h, em que a largura média da faixa críticas era de 3 m. No sentido norte ocorreram exclusivamente em trechos retos, enquanto que no sentido sul em trechos retos ou curvos, Figura K.13 (h).

Na comunidade C_3 , ilustrada na Figura 4.40, verifica-se que a variação média de AUC é de 0,99 para o conjunto de observações.

Figura 4.40 - Variação da AUC para a C_3 considerando a base de severidade com VNF para o período de “depois” (2012-2016).

Na comunidade C_3 com VNF para o período de “depois”, Figura K.14 do Apêndice K (em CD-ROM), os acidentes no sentido norte se concentraram no km 134, km 135, km 139 e km 142, enquanto que no sentido sul no km 135, km 139 e km 141, Figura K.14 (a). Nestes marcos quilométricos foram verificados os maiores VDMs de veículos, Figura K.14 (b).

Em ambos os sentidos da rodovia os tipos de acidentes mais frequentes foram choque, colisão traseira e outros, em trechos mistos e com barreiras, Figura K.14 (c). Os acidentes foram observados em condição do tempo bom, condição de pista seca, condição de visibilidade boa ou parcial, Figura K.14 (d).

Os acidentes da C_3 com VNF para o período de “depois” no sentido norte ocorreram predominantemente em trecho em nível (inclinação entre -2% a 2%) e no trecho sul em trechos em declive (inclinação menor que -2%), Figura 4.54 (e). Os acidentes ocorreram predominantemente com a largura do acostamento de 13 m, Figura 4.54 (f), e com a largura média do acostamento de 3 m, Figura 4.54 (g).

Estes acidentes ocorreram exclusivamente em trechos com 2 faixas, com limite de velocidade de 90 km/h e com a largura média da faixa crítica de 3 m. Em ambos os sentidos da rodovia ocorreram em trechos retos ou curvos, Figura 4.54 (h).

Na comunidade C_4 , ilustrada na Figura 4.41, verifica-se que a variação média de AUC é de 0,99 para o conjunto de observações.

Figura 4.41 - Variação da AUC para a C_4 considerando a base de severidade com VNF para o período de “depois” (2012-2016).

Na comunidade C_4 com VNF para o período de “depois”, Figura K.15 do Apêndice K (em CD-ROM), no sentido norte os acidentes foram mais frequentes nos marcos quilométricos km 134, km 137, km 139, km 140, km 142 e km 146, enquanto que no sentido sul nos marcos quilométricos km 135, km 137 e km 139, Figura K.15 (a). Nestes marcos quilométricos foram verificados os maiores VDMs de veículos, Figura K.15 (b).

No sentido norte os acidentes ocorreram em trechos sem barreiras, com barreiras ou mistos, enquanto no sentido sul em trechos com barreiras ou mistos. Os tipos de acidentes mais frequentes em ambos os sentidos foram colisão traseira, choque, capotamento e outros, Figura K.15 (c). Estes acidentes foram observados predominantemente em condição de tempo bom, condição de pista seca, e condição de visibilidade boa ou parcial, Figura K.15 (d).

E ambos os sentidos da rodovia os acidentes da comunidade C_4 com VNF para o período de “depois” ocorreram com maior frequência em trechos em nível (inclinação de -2% a 2%) e em declive (inclinação menor que -2%), Figura K.15 (e).

Os acidentes foram observados com maior frequência em trechos em que a largura do canteiro central era de 11 m, Figura K.15 (f), e com a largura do acostamento de 3 m, Figura K.15 (g).

Estes acidentes ocorreram exclusivamente em trechos com 2 faixas e com limite de velocidade de 90 km/h, com a largura média da faixa crítica de 3 m. Em ambos os sentidos da rodovia ocorreram em trechos predominantemente retos, Figura K.15 (h).

Acidentes com vítimas fatais no período de antes

Na comunidade C_1 , ilustrada na Figura K.16 do Apêndice K (em CD-ROM), verifica-se que a variação média de AUC é de 0,99 para o conjunto de observações.

Na comunidade C_1 com VF no período de “antes” em ambos os sentidos da rodovia há uma concentração constante de acidentes do km 128 ao km 132, e uma alta concentração no km 134, Figura K.16 (a). Os VDMs de veículos observados neste período são apresentados na Figura K.16 (b). No sentido norte os tipos de acidentes mais frequentes foram atropelamento, tombamento, colisão traseira, colisão transversal e engavetamento, enquanto no sentido sul acidentes do tipo atropelamento, em trechos sem barreiras, com barreiras ou mistos, Figura K.16 (c).

Os acidentes ocorreram predominantemente em condição de tempo bom ou chuva, condição de pista seca ou molhada, e condição de visibilidade parcial, Figura K.16 (d). No sentido norte os acidentes da C_1 com VF no período de “antes” ocorreram predominantemente em nível (inclinação de -2% a 2%), enquanto que no sentido sul em declive (inclinação menor que -2%), Figura K.16 (e). Em geral, ocorreram em trechos com a largura do canteiro central de 11 m, Figura K.16 (f), e a largura média do acostamento de 3 m, Figura K.16 (g).

No sentido norte os acidentes da C_1 com VF no período de “antes” ocorreram em trechos com 3 faixas. No sentido norte os acidentes foram mais frequentes em trechos com limite de velocidade variado, de 80 km/h a 110 km/h, com a largura média da faixa crítica de 3 m, Figura 4.57 (h). No sentido sul os acidentes ocorreram predominantemente em trechos com velocidade entre 100 km/h a 110 km/h, e velocidade da largura média da faixa crítica de 3 m. Ocorreram no sentido norte em trechos com traçado reto ou curvo, enquanto no sentido sul com traçado exclusivamente reto, Figura 4.57 (h).

Na comunidade C_2 com VF no período de “antes”, Figura K.17 do Apêndice K (em CD-ROM), no sentido norte ocorreu alta concentração de acidentes no km 140 e km 141, enquanto que no sentido sul no km 141 e km 142, Figura K.17 (a). Os VDMs de veículos são ilustrados na Figura K.17 (b). Em ambos os sentidos da rodovia os acidentes ocorreram em trechos sem barreiras ou mistos. No sentido norte os tipos de acidentes mais frequentes foram choque e atropelamento, enquanto no sentido atropelamento, choque, colisão traseira, Figura K.17 (c).

Os acidentes da comunidade C_2 com VF no período de “antes” no sentido norte foram observados exclusivamente em condição de tempo bom, condição de pista seca, e condição de visibilidade boa ou parcial. Já no sentido sul, em condição de tempo bom ou com chuva, condição de pista seca ou molhada, e condição de visibilidade boa ou parcial,

Figura K.17 (d). No sentido norte foram observados exclusivamente em trechos em nível (inclinação de 0,50% a 2,0%), enquanto no sentido sul em trechos em nível (inclinação de 0% a 1,90%) ou em declive (inclinação de 4,30%), Figura K.17 (e).

No sentido norte os acidentes da comunidade C_2 com VF no período de “antes” ocorreram em trechos com a largura do canteiro central de 13 m, enquanto que no sentido sul com a largura do canteiro central de 10,90 m a 13,60 m, Figura K.17 (f). No sentido norte ocorreram em trechos com a largura média do acostamento de 3,20 m, enquanto no sentido sul de aproximadamente 3 m. No sentido norte ocorreram exclusivamente em trechos curvos, enquanto no sentido sul em trechos retos ou curvos, Figura K.17 (g).

Em ambos os sentidos da rodovia os acidentes da comunidade C_2 com VF no período de “antes” foram observados em trechos com 3 faixas e com limite de velocidade de 100 km/h, com a largura média da faixa crítica de aproximadamente 3 m, Figura K.17 (h).

Na comunidade C_3 com VF no período de “antes”, Figura K.18 do Apêndice K (em CD-ROM), os acidentes foram observados do km 127 ao km 143, enquanto que a partir do km 144 há um aumento significativo na frequência de acidentes, Figura K.18 (a). Nestes marcos quilométricos foram verificados os maiores VDMs de veículos, Figura K.18 (b).

Estes acidentes ocorreram exclusivamente no sentido norte, em trechos sem ou com barreiras, sendo os tipos de acidentes mais frequentes colisão traseira, colisão frontal, colisão lateral, choque, engavetamento e outros, Figura K.18 (c). Ocorreram em condição de tempo bom ou com chuva, condição de pista seca ou molhada, e condição de visibilidade boa, parcial ou ruim, Figura K.18 (d). Foram verificados em trechos em nível (inclinação de -2% a 2%), em aclive (inclinação superior a 2%) ou declive (inclinação inferior a -2%), Figura K.18 (e).

Os acidentes na C_3 com VF no período de “antes” foram mais frequentes em trechos com largura do canteiro central de 13 m, Figura K.18 (f), e com a largura média do acostamento de 3 m, Figura K.18 (g). Estes acidentes ocorreram exclusivamente em trechos com 3 faixas, com limite de velocidade de 80 km/h a 110 km/h, e com a largura média da faixa crítica de 3,50 m, com traçado reto, Figura K.18 (h).

Acidentes com vítimas fatais no período de depois

O conjunto de observações das comunidades C_1 , C_2 , C_3 e C_4 com VF no período de “antes” apresentaram todas $AUC=1$. Na Figura K.19 do Apêndice K (em CD-ROM), são apresentadas as variáveis que mais contribuíram a ocorrência dos acidentes na comunidade C_1 considerando VF no período de “depois”.

Na comunidade C_1 considerando VF no período de “depois” os acidentes ocorreram exclusivamente no sentido norte da rodovia, havendo uma alta concentração de acidentes no km 128, Figura K.19 (a). Neste marco quilométrico é verificado o maior VDM de veículo para o período observado Figura K.19 (b). Os acidentes ocorreram em trechos com barreiras ou mistos. Os tipos de acidentes mais frequentes foram atropelamento, capotamento, choque, colisão traseira e colisão lateral, Figura K.19 (c). Foram observados predominantemente em condição de tempo bom, condição da pista seca e condição de visibilidade parcial Figura K.19 (d).

Os acidentes da C_1 considerando VF no período de “depois” ocorreram em trechos em nível (inclinação de -2% a 2%), aclive (inclinação maior que 2%) ou declive (inclinação menor que -2%), Figura K.19 (e). Foram observados com maior frequência em trechos com a largura média do canteiro central de 11 m, Figura K.19 (f), e com a largura do acostamento de 3 m, Figura K.19 (g). Os acidentes na C_1 considerando VF no período de “depois” ocorreram exclusivamente em trechos com 2 faixas, com limite de velocidade de 90 km/h a 110 km/h, e com a largura da faixa crítica de 3 m, com traçado reto, Figura K.19 (h).

Na comunidade C_2 com VF no período de “depois” os acidentes foram observados com maior frequência no km 137, sendo constante do km 138 ao km 146, enquanto que no sentido sul ocorreram predominantemente no km 143, Figura K.20 (a). Os VDMs de veículos são representados na Figura K.20 (b). Foram observados em trechos sem barreiras ou com barreiras, sendo os tipos de acidentes mais frequentes no sentido norte colisão traseira, atropelamento, choque e outros, enquanto no sentido sul acidentes do tipo choque e colisão traseira, Figura K.20 (c).

Os acidentes da comunidade C_2 com VF no período de “depois” ocorreram em condição de tempo bom e condição de pista seca. No sentido norte ocorreram tanto em condição de visibilidade boa ou parcial, enquanto que no sentido sul em condição de

visibilidade boa, Figura K.20 (d). Foram observados no sentido norte em trechos em nível (inclinação entre -2% a 2%), aclive (inclinação maior que 2%) ou declive (inclinação inferior a -2%), Figura K.20 (e).

Estes acidentes da C_2 com VF no período de “depois” em ambos os sentidos foram observados com maior frequência em trechos com a largura média do canteiro central de 13 m, Figura K.20 (f), e com a largura do acostamento de 3 m, Figura K.20 (g). Em ambos os sentidos da rodovia ocorreram em trechos com 2 faixas, no sentido norte, com limite de velocidade de 80 km/h a 90 km/h, com a largura média da faixa crítica de 3 m. Enquanto que no sentido sul ocorreram predominantemente em trechos em que o limite de velocidade era de 90 km/h e em trechos com a largura média da faixa crítica de 3 m, com traçado reto, Figura K.20 (h).

Na comunidade C_3 com VF no período de “depois” no sentido norte há uma alta frequência de acidentes no km 134 e km 135, e depois uma constante frequência de acidentes do km 139 ao km 142, enquanto que no sentido sul os acidentes foram constantes no km 134 ao km 135, Figura K.21 (a). Os VDMs de veículos podem ser verificados na Figura K.21 (b).

No sentido norte os acidentes da C_3 com VF no período de “depois” ocorreram em trechos com barreiras ou mistos, sendo os tipos de acidentes mais frequentes por atropelamento e colisão transversal. Enquanto no sentido sul os acidentes ocorreram exclusivamente em trechos com barreiras, sendo os tipos mais frequentes por atropelamento e colisão transversal, Figura K.21 (c).

Os acidentes da C_3 com VF no período de “depois” no sentido norte foram observados em condição de tempo bom, condição de pista seca, e condição de visibilidade boa ou parcial. Enquanto no sentido sul, os acidentes foram observados em condição de tempo bom, condição de pista seca, e condição de visibilidade boa ou parcial, Figura K.21 (d). No sentido norte ocorreram predominantemente em nível (inclinação entre -2% a 2%) e declive (inclinação menor que -2%), e no sentido sul ocorreram predominantemente em trechos em nível com inclinação igual a 1,70%, Figura K.21 (e).

Em ambos os sentidos da rodovia os acidentes com VF no período de “depois” na C_3 ocorreram predominantemente em trechos com a largura média do canteiro central de 11 m, Figura K.21 (f), e com a largura do acostamento de 3 m, Figura K.21 (g). Tanto no sentido norte quanto no sentido sul, os acidentes foram observados exclusivamente em

trechos com 2 faixas, com limite de velocidade de 90 km/h e com a largura da faixa crítica de 3 m. No sentido norte ocorreram com maior frequência em trechos retos ou em curva, enquanto que no sentido sul em trechos curvos, traçado reto, Figura K.21 (h).

As variáveis que contribuíram aos acidentes da comunidade C_4 com VF no período de “depois” são apresentadas na Figura K.22 do Apêndice K (em CD-ROM).

Na comunidade C_4 com VF no período de “depois” os acidentes no sentido norte foram frequentes do km 139 ao km 143, havendo alta concentração no km 144, enquanto que no sentido sul os acidentes são constantes ao longo do km 137 ao km 145, Figura K.22 (a). Os VDMs de veículos são observados na Figura K.22 (b). Em ambos os sentidos da rodovia ocorreram em trechos sem barreiras ou com barreiras. Os tipos de acidentes mais frequentes no sentido norte foram colisão traseira e choque, enquanto no sentido sul colisão traseira, colisão frontal e choque, Figura K.22 (c).

Os acidentes observados na comunidade C_4 com VF no período de “depois” ocorreram exclusivamente em condição de tempo bom e condição de pista seca. No sentido norte, foram observados com maior frequência em condição de visibilidade boa, enquanto que no sentido sul em condição de visibilidade parcial, Figura K.22 (d).

Na comunidade C_4 com VF no período de “depois” no sentido norte ocorreram e trechos em nível (inclinação de até 1,10%) ou em aclave (inclinação de até 4,60%). No sentido sul ocorreram em aclave (inclinação de até 4,50%) ou em declive (inclinação de até 3,10%), Figura K.22 (e). Em ambos os sentidos da rodovia foram ocorreram com largura do canteiro central de 13 m, Figura K.22 (f), e largura do acostamento de 3 m, Figura K.22 (g).

Os acidentes da C_4 com VF no período de “depois” ocorreram exclusivamente em trechos com 2 faixas e com limite de velocidade de 80 km/h. No sentido norte ocorreram em trechos com a largura da faixa crítica variando de 3 m a 3,60 m, enquanto que no sentido sul com largura da faixa crítica variando de 2,50 m a 3 m. Em ambos os sentidos da rodovia ocorreram com traçado reto, Figura K.22 (h).

4.7 Modelagem e predição de acidentes no formato tabela atributo-valor utilizando a abordagem de redes complexas bipartidas

Nesta etapa, foram consideradas tanto no processo de modelagem da rede quanto na PL duas abordagens distintas. A primeira, denominada de “redes geral”, visou identificar o impacto das variáveis selecionadas na seção 4.5. Na segunda, denominada de “redes

simplificada", foram utilizadas exclusivamente as variáveis do modelo HSM, sendo estas: a largura da pista, a largura do acostamento, a largura do canteiro central, a iluminação, o radar, o VDMA e o número de acidentes. Em ambas as abordagens os dados foram explorados considerando exclusivamente os “acidentes totais” e os “acidentes fatais e feridos”, como recomendado no manual do HSM (AASHTO, 2010).

No processo de PL o número de acidentes em uma determinada época foi estimado com base nos dados da época antecessora, com a finalidade de identificar a variação dos pontos críticos de acidentes no tempo e no espaço. O número de acidentes mais prováveis corresponde aos links com escore maior que zero e que são observados quando não se verifica a variação significativa na acurácia da predição fornecida pela métrica *AUC*. Graficamente, verifica-se a variação decrescente da *AUC* e um “cotovelo” no gráfico, a partir do qual as predições se tornam constantes.

Desta forma, a partir da base georreferenciada de acidentes do ano de 2009 foi predito o número de acidentes mais prováveis para o ano de 2010, considerando as duas abordagens propostas. As Figuras 4.42 (a) e 4.42 (b) ilustram, respectivamente, o número de “acidentes totais” e as Figuras 4.43 (a) e 4.43 (b) os “acidentes fatais e feridos”, preditos usando a abordagem de “redes geral” e “redes simplificada” para o ano de 2010.

Figura 4.42 - Predição do número de “acidentes totais” para o ano de 2010.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

Figura 4.43 - Predição do número de “acidentes fatais e feridos” para o ano de 2010.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

Para o ano de 2010 foram preditos com a modelagem “redes geral” 837 acidentes “totais”, Figura 4.42 (a), e 208 acidentes “fatais e feridos”, Figura 4.43 (a), com AUC médio de 0,39 e 0,30, respectivamente. A predição dos acidentes pela modelagem de “redes geral” resultou em uma acurácia e precisão da ordem de 94% e 41%, respectivamente, para os acidentes “totais”, enquanto para os acidentes “fatais e feridos” uma acurácia e precisão da ordem de, respectivamente, 42% e 29%.

Pela modelagem “redes simplificada” foram preditos 802 acidentes “totais”, Figura 4.42 (b), e 193 acidentes “fatais e feridos”, Figura 4.43 (b), com AUC médio de 0,92 e 0,48, respectivamente. Esta modelagem resultou em uma acurácia e precisão da ordem de, respectivamente, 90% e 47% na predição dos acidentes “totais”, e de 38% e 28% na predição dos acidentes “fatais e feridos”. A modelagem “redes simplificada” que utiliza as variáveis selecionadas do HSM apresenta links preditos com maior escore (maior variação do AUC), bem como uma melhor acurácia e precisão dos acidentes “totais” e “fatais e feridos”, quando comparada a modelagem de “redes geral”, o que indica que as variáveis recomendadas pelo HSM são mais explicativas na predição de acidentes, utilizando a modelagem de dados em estruturas de redes complexas bipartidas.

Com a finalidade de comparar as diferentes modelagens de predição de acidentes baseadas em redes complexas bipartidas, foram elaborados testes gráficos baseados no acúmulo residual denominados de “*cure plot*”. Nestes testes os valores dos resíduos acumulados são obtidos pela diferença entre os valores observados (acidentes reais) e os valores preditos de acidentes obtidos pelas modelagens de predição propostas (redes “geral” e “redes simplificada”), expressos em relação ao volume de tráfego (VDMA), (Hauer e Bamfo, 1997; Moghaddam *et al.*, 2017; Young e Park, 2013).

Os resíduos acumulados abaixo de zero indicam que o modelo proposto subestima (*underfitting*) o número de acidentes “totais” ou “fatais e feridos” preditos, enquanto os

resíduos acumulados acima de zero indicam que o modelo superestima (*overfitting*) o número de acidentes “totais” ou “fatais e feridos” preditos. Os modelos que melhor se ajustam aos dados são verificados quando os resíduos acumulados oscilam no intervalo de dois desvios padrão ($\pm 2\sigma$), para mais ou para menos (com limite de confiança de 95%), (Hauer, 2015). Quanto mais a curva dos resíduos acumulados se aproxima do eixo horizontal, ou seja, quanto menor a oscilação dos resíduos em torno deste eixo melhor será o modelo selecionado (Hauer e Bamfo, 1997; Moghaddam *et al.*, 2017).

Foram construídos 8 gráficos pelo teste “*cure plot*” para os anos de 2010 a 2012 e de 2012 a 2016 (Figura 4.44, 4.47, 4.50, 4.53, 4.56, 4.59, 4.62 e 4.65), antes e após a intervenção da concessionária no trecho em estudo. Nestes gráficos, as linhas azuis representam os resíduos acumulados na predição dos acidentes “totais”, as linhas vermelhas os resíduos acumulados obtidos na predição dos acidentes “fatais e feridos”, enquanto as linhas pontilhadas os limites positivos e negativos de dois desvios padrão, respectivamente, acima e abaixo de zero.

As Figuras 4.44 (a) e 4.44 (b) ilustram os gráficos dos resíduos acumulados na predição dos acidentes “totais”, enquanto que as Figuras 4.44 (c) e 4.44 (d) ilustram os resíduos acumulados obtidos na predição dos acidentes “fatais e feridos”, pelas abordagens de predição por “redes geral” e “redes simplificada” para o ano de 2010.

Figura 4.44 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2010 utilizando diferentes modelagens.

Verifica-se por meio dos gráficos “*cure plot*” ilustrados nas Figura 4.44 (a) a 4.44 (d) que os resíduos acumulados obtidos pelas modelagens de “redes geral” e “redes simplificadas” para o ano de 2010 oscilam de forma admissível entre os limites de $\pm 2\sigma$, e se afastam do eixo horizontal em relação ao limite inferior para VDMs maiores ou igual a 55250 veículos, o que indica que a partir deste ponto pode ocorrer uma subestimação no número de acidentes.

As Figuras 4.45 (a) e 4.45 (b) ilustram o número de “acidentes totais” e as Figuras 4.46 (a) e 4.46 (b) os “acidentes fatais e feridos”, respectivamente, usando a abordagem de “redes geral” e “redes simplificada” para o ano de 2011.

Figura 4.45 - Predição do número de “acidentes totais” para o ano de 2011.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

Figura 4.46 - Predição do número de “acidentes fatais e feridos” para o ano de 2011.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

Para o ano de 2011 foram preditos com a modelagem “redes geral” 843 acidentes “totais”, Figura 4.45 (a), e 219 acidentes “fatais e feridos”, Figura 4.46 (a), com *AUC* médio de 0,51 e 0,44, respectivamente. A predição dos acidentes pela modelagem de “redes geral” resultou em uma acurácia e precisão da ordem de 74% e 42%, respectivamente, para os acidentes “totais”, enquanto para os acidentes “fatais e feridos” uma acurácia e precisão da ordem de, respectivamente, 60% e 50%.

Pela modelagem “redes simplificada” para o ano de 2011 foram preditos 839 acidentes “totais”, Figura 4.45 (b), e 217 acidentes “fatais e feridos”, Figura 4.46 (b), com AUC médio de 0,73 e 0,74, respectivamente. Esta modelagem resultou em uma acurácia e precisão da ordem de, respectivamente, 73% e 42% na predição dos acidentes “totais”, e de 35% e 26% na predição dos acidentes “fatais e feridos”.

Para o ano de 2011 os gráficos “*cure plot*” ilustrados nas Figura 4.47 (a) a 4.47 (d) indicaram que os resíduos acumulados obtidos pelas modelagens de “redes geral” e “redes simplificadas” para o ano de 2011 oscilaram de consideravelmente em relação ao limite inferior para VDMs maiores que 55250 veículos, o que indica que a partir deste ponto pode ocorrer uma subestimação no número de acidentes em ambas as modelagens.

Figura 4.47 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2011 utilizando diferentes modelagens.

As Figuras 4.48 (a) e 4.48 (b) ilustram, respectivamente, o número de “acidentes totais” preditos e as Figuras 4.49 (a) e 4.49 (b) os “acidentes fatais e feridos”, usando a abordagem de “redes geral” e “redes simplificada” para o ano de 2012 no período de “antes”.

Figura 4.48 - Predição do número de “acidentes totais” para o ano de 2012 no período de “antes”.

Figura 4.49 - Predição do número de “acidentes fatais e feridos” para o ano de 2012 no período de “antes”.

Para o ano de 2012 no período de “antes” foram preditos com a modelagem “redes geral” 584 acidentes “totais”, Figura 4.48 (a), e 126 acidentes “fatais e feridos”, Figura 4.49 (a), com AUC médio de 0,27 e 0,70, respectivamente. A predição dos acidentes pela modelagem de “redes geral” resultou em uma acurácia e precisão da ordem de 58% e 37%, respectivamente, para os acidentes “totais”, enquanto para os acidentes “fatais e feridos” uma acurácia e precisão da ordem de, respectivamente, 19% e 16%.

Pela modelagem “redes simplificada” para o ano de 2012 antes foram preditos 591 acidentes “totais”, Figura 4.48 (b), e 131 acidentes “fatais e feridos”, Figura 4.49 (b), com AUC médio de 0,21 e 0,70, respectivamente. Esta modelagem resultou em uma acurácia e precisão da ordem de, respectivamente, 62% e 38% na predição dos acidentes “totais”, e de 20% e 17% na predição dos acidentes “fatais e feridos”.

Para o ano de 2012 no período de antes os gráficos “*cure plot*” ilustrados nas Figura 4.50 (a) a 4.50 (d) indicam que os resíduos acumulados obtidos pelas modelagens de “redes geral” e “redes simplificadas”. A predição dos acidentes “totais” usando a modelagem de “redes geral”, Figura 4.50 (a), oscilou consideravelmente para volumes de tráfego maiores que 21.350 veículos. Para as demais predições os resíduos acumulados

oscilaram de forma admissível entre os limites de $\pm 2\sigma$, conforme ilustram as Figuras 4.50 (b) a 4.50 (d), embora também nestes casos seja observado um deslocamento dos resíduos acumulados em direção ao limite inferior para VDMs superiores a 55.250 veículos.

Na modelagem dos acidentes “fatais e feridos”, Figuras 4.50 (c) e 4.50 (d), verifica-se uma admissível oscilação dos resíduos acumulados em direção ao limite superior para trechos com volume de tráfego entre 21.350 a 55.250, o que indica superestimação de acidentes nestes trechos.

Figura 4.50 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2012 (antes) utilizando diferentes modelagens.

As Figuras 4.51 (a) e 4.51 (b) ilustram, respectivamente, o número de “acidentes totais” preditos e as Figuras 4.52 (a) e 4.52 (b) os “acidentes fatais e feridos”, usando a abordagem de “redes geral” e “redes simplificada” para o ano de 2012 no período de depois”.

Figura 4.51 - Predição do número de “acidentes totais” para o ano de 2012 no período de “depois”.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

Figura 4.52 - Predição do número de “acidentes fatais e feridos” para o ano de 2012 no período de “depois”.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

Para o ano de 2012 no período de “depois” foram preditos com a modelagem “redes geral” 278 acidentes “totais”, Figura 4.51 (a), e 80 acidentes “fatais e feridos”, Figura 4.52 (a), com AUC médio de 0,33 e 0,38, respectivamente. A predição dos acidentes pela modelagem de “redes geral” resultou em uma acurácia e precisão da ordem de 30% e 23%, respectivamente, para os acidentes “totais”, enquanto para os acidentes “fatais e feridos” uma acurácia e precisão da ordem de, respectivamente, 42% e 30%.

Pela modelagem “redes simplificada” para o ano de 2012 “depois” foram preditos 277 acidentes “totais”, Figura 4.51 (b), e 80 acidentes “fatais e feridos”, Figura 4.52 (b), com AUC médio de 0,14 e 0,38, respectivamente. Esta modelagem resultou em uma acurácia e precisão da ordem de, respectivamente, 30% e 23% na predição dos acidentes “totais”, assim como na modelagem por “redes geral”, e de 42% e 30% na predição dos acidentes “fatais e feridos”.

As Figuras 4.53 (a) a 4.53 (d) ilustram os gráficos “*cure plot*” obtidos pelas modelagens de “redes geral” e “redes simplificadas” para o ano de 2012 no período de depois. A predição dos acidentes “totais” usando a modelagem de “redes geral”, Figura

4.53 (a), oscilou consideravelmente para volumes de tráfego maiores que 55.250 veículos. Para as demais previsões os resíduos acumulados oscilaram de forma admissível entre os limites de $\pm 2\sigma$, conforme ilustram as Figuras 4.53 (b) a 4.53 (d), excedendo os limites estabelecidos para VDMs superiores a 57.050 veículos.

Figura 4.53 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2012 (depois) utilizando diferentes modelagens.

Para o ano de 2013 foram preditos com a modelagem “redes geral” 743 acidentes “totais”, Figura 4.54 (a), e 200 acidentes “fatais e feridos”, Figura 4.55 (a), com AUC médio de 0,37 e 0,27, respectivamente. A predição dos acidentes pela modelagem de “redes geral” resultou em uma acurácia e precisão da ordem de 74% e 43%, respectivamente, para os acidentes “totais”, enquanto para os acidentes “fatais e feridos” uma acurácia e precisão da ordem de, respectivamente, 29% e 22%.

Considerando a modelagem de “redes simplificada” para o ano de 2013 foram preditos 781 acidentes “totais”, Figura 4.54 (b), e 199 acidentes “fatais e feridos”, Figura 4.55 (b), com AUC médio de 0,42 e 0,39, respectivamente. Esta modelagem resultou em uma acurácia e precisão da ordem de, respectivamente, 78% e 44% na predição dos acidentes “totais”, assim como na modelagem por “redes geral”, e de 28% e 22% na predição dos acidentes “fatais e feridos”.

Figura 4.54 - Predição do número de “acidentes totais” para o ano de 2013.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

Figura 4.55 - Predição do número de “acidentes fatais e feridos” para o ano de 2013.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

As Figuras 4.56 (a) a 4.56 (d) ilustram os gráficos “*cure plot*” obtidos pelas modelagens de “redes geral” e “redes simplificadas” para o ano de 2013.

Figura 4.56 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2013 utilizando diferentes modelagens.

(a) Acidentes Totais: “Redes Geral”.

(b) Acidentes Totais: “Redes Simplificada”.

(c) Acidentes Fatais e Feridos: “Redes Geral”.

(d) Acidentes Fatais e Feridos: “Redes Simplificada”.

Para o ano de 2013, a predição dos acidentes “totais” usando a modelagem de “redes geral”, Figura 4.56 (a), oscilou consideravelmente para volumes de tráfego maiores que 55.250 veículos. Para as demais predições os resíduos acumulados oscilaram de forma admissível entre $\pm 2\sigma$, conforme ilustram as Figuras 4.56 (b) a 4.56 (d), excedendo os limites estabelecidos para VDMs superiores a 57.050 veículos.

Para o ano de 2014 foram preditos com a modelagem “redes geral” 625 acidentes “totais”, Figura 4.57 (a), e 191 acidentes “fatais e feridos”, Figura 4.58 (a), com *AUC* médio de 0,15 e 0,10, respectivamente. A predição dos acidentes pela modelagem de “redes geral” resultou em uma acurácia e precisão da ordem de 63% e 38%, respectivamente, para os acidentes “totais”, enquanto para os acidentes “fatais e feridos” uma acurácia e precisão da ordem de, respectivamente, 36% e 26%.

A modelagem de “redes simplificada” para o ano de 2014 resultou na predição de 637 acidentes “totais”, Figura 4.57 (b), e 185 acidentes “fatais e feridos”, Figura 4.58 (b), com *AUC* médio de 0,12 e 0,14, respectivamente. Esta modelagem resultou em uma acurácia e precisão da ordem de, respectivamente, 64% e 39% na predição dos acidentes “totais”, assim como na modelagem por “redes geral”, e de 34% e 26% na predição dos acidentes “fatais e feridos”.

Figura 4.57 - Predição do número de “acidentes totais” para o ano de 2014.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

Figura 4.58 - Predição do número de “acidentes fatais e feridos” para o ano de 2014.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

As Figuras 4.59 (a) a 4.59 (d) ilustram os gráficos “*cure plot*” obtidos pelas modelagens de “redes geral” e “redes simplificadas” para o ano de 2014. Ambas as modelagens para o ano de 2014, oscilaram consideravelmente para volumes de tráfego maiores que 55.250 veículos, excedendo nestes casos o limite inferior estabelecido.

Figura 4.59 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2014 utilizando diferentes modelagens.

Para o ano de 2015 foram preditos com a modelagem “redes geral” 635 acidentes “totais”, Figura 4.60 (a), e 156 acidentes “fatais e feridos”, Figura 4.61 (a), com *AUC* médio de 0,22 e 0,31, respectivamente. A predição dos acidentes pela modelagem de “redes geral” resultou em uma acurácia e precisão da ordem de 56% e 36%, respectivamente, para os acidentes “totais”, enquanto para os acidentes “fatais e feridos” uma acurácia e precisão da ordem de, respectivamente, 25% e 20%.

A modelagem de “redes simplificada” para o ano de 2015 resultou na predição de 529 acidentes “totais”, Figura 4.60 (b), e 158 acidentes “fatais e feridos”, Figura 4.61 (b), com *AUC* médio de 0,23 e 0,52, respectivamente. Esta modelagem resultou em uma acurácia e precisão da ordem de, respectivamente, 53% e 25% na predição dos acidentes “totais”, assim como na modelagem por “redes geral”, e de 25% e 20% na predição dos acidentes “fatais e feridos”.

Figura 4.60 - Predição do número de “acidentes totais” para o ano de 2015.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

Figura 4.61 - Predição do número de “acidentes fatais e feridos” para o ano de 2015.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

As Figuras 4.62 (a) a 4.62 (d) ilustram os gráficos “*cure plot*” obtidos pelas modelagens de “redes geral” e “redes simplificadas” para o ano de 2015. A predição dos acidentes “totais” por ambas as modelagens para o ano de 2015 oscilaram consideravelmente para volumes de tráfego superiores a 85.000 veículos, excedendo nestes casos o limite inferior estabelecido, conforme ilustram as Figuras 4.62 (a) e 4.62 (b). Enquanto que na predição dos acidentes “fatais e feridos”, por ambas as modelagens, os resíduos acumulados oscilaram de forma admissível entre $\pm 2\sigma$, conforme ilustram as Figuras 4.62 (c) a 4.62 (d), excedendo os limites estabelecidos para VDMs superiores a 57.050 veículos.

Figura 4.62 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2015 utilizando diferentes modelagens.

Para o ano de 2016 foram preditos com a modelagem “redes geral” 380 acidentes “totais”, Figura 4.63 (a), e 119 acidentes “fatais e feridos”, Figura 4.64 (a), com *AUC* médio de 0,34 e 0,38, respectivamente. A predição dos acidentes pela modelagem de “redes geral” resultou em uma acurácia e precisão da ordem de 38% e 28%, respectivamente, para os acidentes “totais”, enquanto para os acidentes “fatais e feridos” uma acurácia e precisão da ordem de, respectivamente, 23% e 19%.

A modelagem de “redes simplificada” para o ano de 2016 resultou na predição de 376 acidentes “totais”, Figura 4.63 (b), e 119 acidentes “fatais e feridos”, Figura 4.64 (b), com *AUC* médio de 0,34 e 0,38, respectivamente. Esta modelagem resultou em uma acurácia e precisão da ordem de, respectivamente, 38% e 27% na predição dos acidentes “totais”, assim como na modelagem por “redes geral”, e de 23% e 19% na predição dos acidentes “fatais e feridos”.

Figura 4.63 - Predição do número de “acidentes totais” para o ano de 2016.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

Figura 4.64 - Predição do número de “acidentes fatais e feridos” para o ano de 2016.

(a) Modelagem “redes geral”.

(b) Modelagem “redes simplificada”.

As Figuras 4.65 (a) a 4.65 (d) ilustram os gráficos “*cure plot*” obtidos pelas modelagens de “redes geral” e “redes simplificadas” para o ano de 2016. Para o ano de 2016, a predição dos acidentes “totais” por ambas as modelagens resultou em resíduos acumulados que oscilaram consideravelmente para volumes de tráfego entre 55250 a 57050 veículos, o que sugere que pode haver uma subestimação de acidentes nestes segmentos, conforme ilustram as Figuras 4.65 (a) e 4.65 (b). A modelagem dos acidentes “fatais e feridos” por ambas as abordagens de predição resultaram em resíduos acumulados que oscilaram de forma admissível entre os limites estabelecidos, conforme ilustram as Figuras 4.65 (b) e 4.65 (d).

Figura 4.65 - Gráficos dos resíduos acumulados obtidos na predição dos “acidentes totais” e “acidentes fatais e feridos” para 2016 utilizando diferentes modelagens.

As Tabelas 4.33 e 4.34 apresentam, respectivamente, a frequência dos acidentes totais e “fatais e feridos” observados (reais), registrados na SP-065 pela concessionária “Rota das Bandeiras”, bem como a frequência dos acidentes obtidas pelas modelagens de predição propostas em estruturas de redes (“redes geral” e “redes simplificada”), para o período antes e depois da concessão da rodovia. Com a finalidade de comparar os resultados obtidos também são apresentadas as frequências de acidentes estimadas pelo modelo do HSM, bem como seus respectivos valores ajustados pelo método EB (AASHTO, 2010).

Tabela 4.33 – Frequência de acidentes “totais” para o período de “antes” e “depois” da concessão da rodovia SP-065.

Período	Ano	Totais				
		Observados	Redes Geral	Redes Simplificada	HSM	EB
Antes	2009	513	-	-	-	-
	2010	791	837	802	836	810
	2011	827	843	839	874	833
	2012	570	584	591	601	576
Depois	2012	262	278	277	256	260
	2013	741	743	781	722	740
	2014	617	625	637	600	614
	2015	514	562	529	501	514
	2016	367	380	376	356	364

Tabela 4.34 – Frequência de acidentes “fatais e feridos” para o período de “antes” e “depois” da concessão da rodovia SP-065.

Período	Ano	Fatais e Feridos				
		Observados	Redes Geral	Redes Simplificada	HSM	EB
Antes	2009	154	-	-	-	-
	2010	180	208	193	190	188
	2011	205	219	217	217	209
	2012	123	126	131	130	125
Depois	2012	77	80	80	75	75
	2013	182	200	199	177	179
	2014	168	191	185	164	166
	2015	150	156	158	146	149
	2016	109	119	119	106	107

A Figura 4.66 apresenta a variação percentual entre os acidentes observados e os acidentes “totais” preditos pelas abordagens de redes complexas bipartidas (“redes geral” e “redes simplificada”), HSM e EB. Como pode-se observar na Figura 4.66 (a) para o período de “antes” a modelagem “redes simplificada” foi a que mais se aproximou dos valores estimados pelo método EB, seguindo as recomendações da AASHTO (2010). Além disso, destaca-se que a modelagem “redes geral” apresentou menores variações percentuais quando comparada aos valores estimados pelo HSM, ou seja, quando os valores deste modelo não ajustados ou calibrados pelo EB.

Figura 4.66 – Variação percentual dos acidentes “totais” utilizando diferentes abordagens de predição para o período de “antes” e “depois” da concessão da SP-065.

(a) Variação percentual na predição dos acidentes “totais” no período de antes.

(b) Variação percentual na predição dos acidentes “totais” no período de depois.

Na Figura 4.66 (b) verifica-se que dentre as modelagens de predição propostas a de “redes simplificada” apresentou as menores variações percentuais na predição dos acidentes “totais” para o período de “depois”. No entanto, o modelo HSM e calibrado pelo EB apresentaram neste caso as menores variações percentuais em relação aos valores

observados. Ressalta-se, porém, que a modelagem “redes simplificada” se aproximou mais do modelo HSM a partir do ano de 2014.

A Figura 4.67 apresenta a variação percentual obtida entre os acidentes “fatais e feridos” observados e preditos pelas modelagens de redes complexas bipartidas (“redes Geral” e “redes simplificada”), HSM e EB.

Figura 4.67 – Variação percentual dos acidentes “fatais e feridos” utilizando diferentes abordagens de predição para o período de “antes” e “depois” da concessão da SP-065.

(a) Variação percentual na predição dos acidentes “fatais e feridos” no período de antes.

(b) Variação percentual na predição dos acidentes “fatais e feridos” no período de depois.

Como pode-se observar na Figura 4.67 (a) para o período de “antes” a modelagem “redes simplificada” converge com a do HSM, embora os valores ajustados pelo EB sejam mais próximos dos valores observados. A modelagem de “redes geral” tende a se aproximar do HSM calibrado pelo EB a partir do ano de 2011, no entanto apresenta variações percentuais mais inconstantes.

Na Figura 4.67 (b) para o período de “depois” as modelagens de dados em redes complexas bipartidas apresentaram tendências de variação semelhantes, embora a modelagem por “redes simplificada” apresente as menores variações percentuais para o conjunto de dados analisado. Neste período, as modelagens de predição propostas foram mais discrepantes em relação ao modelo HSM e de seus valores calibrados pelo EB (AASHTO, 2010). A análise da variação percentual sugere que para o período de “depois” os dados podem ter estruturas ou padrões mais anômalos, em razão deste período ser afetado pela intervenção ou obras realizadas pela concessionária no trecho em estudo.

4.8 Análise espacial das modelagens de predição de acidentes

Nesta seção são apresentadas as espacializações e análises dos resultados obtidos por meio das diferentes modelagens de predição de acidentes abordadas nesta pesquisa. É possível por meio de representações espaciais como os mapas de calor (“*heat maps*”), visualizar e identificar variações espaço-temporais dos pontos críticos de acidentes, em cada quilômetro do trecho da rodovia em estudo.

Nos *heat maps*, a frequência ou densidade de acidentes preditos é representada em uma escala de cores. O gradiente de cor utilizado na elaboração deste tipo de mapa segue um padrão hierárquico, variando entre valores mínimos e máximos. Em geral, os valores mínimos são representados na cor azul escuro, enquanto as cores quentes representam pontos com elevada densidade ou frequência de acidentes preditos, em cada marco quilométrico do trecho analisado, com gradiente máximo na coloração vermelho escuro.

A Figura 4.90 apresenta a espacialização dos acidentes reais registrados no ano de 2009 no trecho compreendido entre o km 125 ao km 145,500 da SP-065.

Ao se analisar microscopicamente os acidentes ocorridos em 2009, Figura 4.68 (a), verifica-se que a maior parte dos acidentes ocorreram em regiões localizadas nas proximidades de dispositivos de acesso ou saída à rodovia. Os acidentes “sem vítimas”, Figura 4.68 (b) aumentam à medida que se aproxima da região metropolitana da cidade de Campinas, ou seja, quanto maior a densidade de conglomerados urbanos maior a concentração de acidentes “sem vítimas”, como se verifica, por exemplo, do km 134 ao km 145,500.

No ano de 2009, os acidentes com “feridos”, Figura 4.68 (c), foram mais recorrentes no km 128 (trevo das “Alças de Souza), km 131 (trevo do Galeria) ao km 134 (trevo de Mogi), km 137 (acesso ao Parque Dom Pedro Shopping), km 139, km 144 e km 145. E os acidentes “fatais”, Figura 4.68 (d), ocorreram predominantemente no km 131, km 132 (trevo do Walmart e próximo a região do Sam’s Club), km 134, km 137, km 139, km 142 e do km 144 ao km 145,500.

Figura 4.68 – Espacialização dos acidentes reais observados no ano de 2009 na rodovia SP-065 do km 125 ao km 145,500.

A partir dos acidentes “reais” observados em 2009 foram analisados espacialmente os acidentes totais “reais” e “fatais e feridos”, em relação aos acidentes preditos no ano de 2010, conforme ilustram as Figuras 4.69 (a) a 4.69 (d) e as Figuras 4.70 (a) a 4.70 (d), respectivamente.

Pode-se verificar para o ano de 2010 tanto para os acidentes “totais”, Figura 4.69, quanto para os acidentes “fatais e feridos”, Figura 4.70, que os *heat maps* obtidos pelas abordagens de predição propostas na pesquisa se aproximam mais da espacialização dos acidentes reais do que a abordagem do HSM, calibrado pelo método EB, conforme recomendação da AASHTO (2010).

Figura 4.69 – Espacialização dos acidentes “totais” para o ano de 2010 na rodovia SP-065 do km 125 ao km 145,500.

Ressalta-se que em relação aos acidentes “fatais e feridos”, Figura 4.70, observados em 2010 apenas no km 126 não foram registrados acidentes. No entanto, todas as abordagens de predição, incluindo a do HSM calibrada pelo EB, estimaram acidentes para esta localização, embora com uma menor frequência.

Em relação aos acidentes “totais” preditos para 2010, Figura 4.69, verifica-se que houve um aumento da ocorrência de acidentes em relação ao ano de 2009, conforme descrito na Tabela 4.33. Esta análise também se estende aos acidentes “fatais e feridos” para o período analisado, Tabela 4.34. Ressalta-se, que todas as abordagens de predição de acidentes utilizadas na pesquisa foram sensíveis a este aumento.

Figura 4.70 – Espacialização dos acidentes “fatais e feridos” para o ano de 2010 na rodovia SP-065 do km 125 ao km 145,500.

A Figura 4.71 apresenta a espacialização dos acidentes “totais” registrados no ano de 2011 no trecho compreendido entre o km 125 ao km 145,500 da SP-065.

Para o ano de 2011 a espacialização dos acidentes “totais” preditos pelas abordagens propostas sugerem uma maior ocorrência de pontos críticos, Figuras 4.71 (b) e 4.71 (c), quando comparada as ocorrências “totais” observadas neste ano, conforme ilustra a Figura 4.71 (a). Em contrapartida, a espacialização pelo modelo HSM calibrado pelo método EB, Figura 4.71 (d), apresentou melhores resultados. Esta análise converge com os valores apresentados pelo teste *cure plot* descrito na seção 4.7, Figura 4.47, em que se verifica que a predição utilizando redes complexas bipartidas foi mais sensível ao aumento do número de veículos. Além disso, conforme mencionado na seção 4.7, houve

um aumento no número de acidentes “totais” e “fatais e feridos” de 2010 a 2011, como ilustrado na Tabela 4.34.

Figura 4.71 – Espacialização dos acidentes “totais” para o ano de 2011 na rodovia SP-065 do km 125 ao km 145,500.

Em relação aos acidentes “fatais e feridos”, Figura 4.72, as espacializações dos acidentes preditos pelas modelagens propostas foram semelhantes. Os pontos críticos que mais culminaram na ocorrência de acidentes “fatais e feridos” no ano de 2011 foram observados nos marcos quilométricos km 128, km 129, km 131 ao km 139 e a partir do km 145. No entanto, o mapeamento da acidentalidade viária utilizando as predições do modelo HSM calibrado pelo método EB foi mais próxima da acidentalidade observada.

Ao se comparar a acidentalidade viária entre os anos de 2010 e 2011 verifica-se que houve um considerado aumento do número de acidentes “totais”, Tabela 4.33, e dos

acidentes “fatais e feridos”, Tabela 4.34. E que esta variação foi detectada pelas diferentes abordagens de predição utilizadas na pesquisa.

Figura 4.72 – Espacialização dos acidentes “fatais e feridos” para o ano de 2011 na rodovia SP-065 do km 125 ao km 145,500.

A Figura 4.73 apresenta a espacialização dos acidentes “totais” registrados no ano de 2012 no período de “antes” no trecho compreendido entre o km 125 ao km 145,500 da SP-065. Pode-se verificar que a abordagem de predição “redes simplificada”, Figura 4.73 (b), apresentou uma melhor espacialização das ocorrências quando comparada a abordagem “redes geral”, o que converge com os resultados de ajustes de modelagem apresentados na seção 4.7 (Figura 4.50). Todavia, a abordagem de predição usando o modelo HSM calibrado pelo EB, Figura 4.50 (d), foi semelhante a espacialização dos

acidentes reais ou observados para o ano de 2012 no período de “antes”, conforme ilustra a Figura 4.73 (a).

Figura 4.73 – Espacialização dos acidentes “totais” para o ano de 2012 no período de “antes” na rodovia SP-065 do km 125 ao km 145,500.

A espacialização dos acidentes “fatais e feridos” para o ano de 2012 no período de “antes” é apresentada na Figura 4.74. Neste cenário, a espacialização das ocorrências utilizando a modelagem “redes simplificada” apresentou melhores resultados quando comparada a “redes geral”, o que também foi verificado no ajuste dos dados (Figura 4.50). No entanto, a acidentalidade viária para 2012 no período de “antes” foi melhor representada e especializada pelo modelo HSM calibrado pelo método EB, conforme ilustra a Figura 4.74 (d).

Figura 4.74 – Espacialização dos acidentes “fatais e feridos” para o ano de 2012 no período de “antes” na rodovia SP-065 do km 125 ao km 145,500.

A Figura 4.75 apresenta a espacialização dos acidentes “totais” registrados no ano de 2012 no período de “depois” no trecho compreendido entre o km 125 ao km 145,500 da SP-065. Neste cenário, a abordagem de “redes simplificada”, assim como no período de 2012 “antes”, apresentou melhor ajuste e espacialização das previsões, conforme ilustra a Figura 4.75 (c), quando comparada a espacialização dos acidentes preditos pela modelagem de “redes geral”, Figura 4.75 (b).

Para o ano de 2012 “antes” e “depois”, Figura 4.75 (a), a modelagem “redes simplificada” se ajustou melhor ao modelo HSM calibrado pelo EB, conforme ilustra a Figura 4.75 (d).

Figura 4.75 – Espacialização dos acidentes “totais” para o ano de 2012 no período de “depois” na rodovia SP-065 do km 125 ao km 145,500.

A espacialização dos acidentes “fatais e feridos” para o ano de 2012 no período de “depois” é apresentada na Figura 4.76. Neste caso, a modelagem por “redes simplificada”, Figura 4.76 (c), resultou em uma melhor espacialização das ocorrências preditas para este nível de severidade, quando comparada aos acidentes “totais” observados, Figura 4.76 (a), bem como as demais modelagens de predição utilizadas nesta pesquisa, conforme ilustram as Figuras 4.76 (b) e 4.76 (d).

Em relação ao ano de 2011 verifica-se que o número de acidentes “totais” e “fatais e feridos” aumentou em 2012. No entanto, a maior parte dos acidentes foram registrados no período que antecede a intervenção da concessionária no trecho em estudo, ou seja, no período de 2012 “antes”. Esta constatação sugere que as intervenções implementadas pela

concessionária visando a redução de acidentes e a modificação da acidentalidade viária no trecho foram eficientes no período de 2012 “depois”.

Figura 4.76 – Espacialização dos acidentes “fatais e feridos” para o ano de 2012 no período de “depois” na rodovia SP-065 do km 125 ao km 145,500.

A Figura 4.77 apresenta a espacialização dos acidentes “totais” registrados no ano de 2013 no trecho compreendido entre o km 125 ao km 145,500 da SP-065. Ao analisarmos as espacializações verifica-se que a abordagem por redes simplificadas, Figura 4.77 (a) apresentou resultados bem próximos à espacialização dos acidentes observados, Figura 4.77 (a) e ao HSM calibrado, Figura 4.77 (d). Em contrapartida, a modelagem por “redes geral” apresentou espacializações de ocorrências mais discrepantes, conforme ilustra a Figura 4.77 (b). Tal análise é análoga aos acidentes preditos “fatais e feridos” para o ano de 2013, conforme ilustra a Figura 4.78.

Figura 4.77 – Espacialização dos acidentes “totais” para o ano de 2013 na rodovia SP-065 do km 125 ao km 145,500.

Ao se analisar a evolução dos acidentes verifica-se que houve uma redução do número de acidentes “totais” entre os anos de 2012 a 2013, Tabela 4.33, assim como para o número de ocorrências envolvendo vítimas “fatais e feridos”, Tabela 4.34. Ressalta-se que todas as modelagens de predição utilizadas nesta pesquisa foram sensíveis a detecção destas variações da acidentalidade, como pode-se observar nas Figuras 4.77 e 4.78.

Figura 4.78 – Espacialização dos acidentes “fatais e feridos” para o ano de 2013 na rodovia SP-065 do km 125 ao km 145,500.

A Figura 4.79 apresenta a espacialização dos acidentes “totais” registrados no ano de 2014 no trecho compreendido entre o km 125 ao km 145,500 da SP-065. Pode-se verificar que ambas as abordagens de predição propostas apresentaram resultados similares entre si, conforme ilustram as Figuras 4.79 (b) e 4.79 (c). Além disso, em razão das modelagens terem sido sensíveis ao aumento do número de veículos, verifica-se que a espacialização dos dados pelo modelo HSM calibrado, Figura 4.79 (a), apresentou resultados mais próximos aos acidentes “totais” observados no ano de 2014, Figura 4.79 (a).

Figura 4.79 – Espacialização dos acidentes “totais” para o ano de 2014 na rodovia SP-065 do km 125 ao km 145,500.

A Figura 4.80 apresenta a espacialização dos acidentes “fatais e feridos” registrados no ano de 2014 no trecho compreendido entre o km 125 ao km 145,500 da SP-065. Verifica-se que a abordagem por “redes simplificada”, Figura 4.80 (c), proporcionou um melhor mapeamento dos pontos críticos quando comparada a abordagem de predição “redes geral”, Figura 4.80 (b). Embora os resultados obtidos tenham sido satisfatórios, ressalta-se que a modelagem foi sensível ao volume de tráfego, Figura 4.59, o que implica que os dados tenham sido mais adequadamente representados pontualmente pelo modelo HSM calibrado pelo EB, conforme ilustra a Figura 4.80 (d).

Considerando a variação da acidentalidade viária entre os anos de 2013 a 2014 verifica-se que houve uma redução do número de acidentes “totais” e “fatais e feridos”, conforme descrito na seção 4.7, respectivamente nas Tabelas 4.33 e 4.34.

Figura 4.80 – Espacialização dos acidentes “fatais e feridos” para o ano de 2014 na rodovia SP-065 do km 125 ao km 145,500.

As Figuras 4.81 e 4.82 apresentam, respectivamente, a espacialização dos acidentes “totais” e “fatais e feridos” para o ano de 2015 no trecho compreendido entre o km 125 ao km 145,500 da SP-065.

Verifica-se por meio dos acidentes “totais” observados no ano de 2015, Figura 4.81 (a), que as abordagens de predição por redes complexas bipartidas, Figura 4.81 (b) e 4.81 (c) foram bem semelhantes, o que se justifica pelo ajuste dos dados a estas modelagens apresentados na Figura 4.62. Todavia, a modelagem de predição dos “acidentes totais” para o ano de 2015 foi mais precisa pela abordagem do HSM calibrado pelo método EB.

Figura 4.81 – Espacialização dos acidentes “totais” para o ano de 2015 na rodovia SP-065 do km 125 ao km 145,500.

Quanto aos acidentes “totais” preditos verifica-se que as abordagens de predição por “redes geral”, Figura 4.82 (b), “redes simplificada”, Figura 4.82 (c), e HSM, Figura 4.82 (d), foram convergentes. Tal fato pode ser verificado pelo adequado ajuste dos dados a abordagem de predição proposta, conforme ilustra a Figura 4.82 (c) e 4.82 (d). Ressalta-se que a modelagem pelo método HSM calibrada pelo EB espacialmente foi mais próxima à espacialização dos acidentes “reais” observados no ano de 2015, Figura 4.82 (a). No entanto, também se observam falhas na predição pelo método HSM, como por exemplo entre os marcos quilométricos km 126 ao km 127, e no km 138.

Em relação ao ano de 2014 verifica-se que em o ano de 2015 houve a redução do número de acidentes “totais” e “fatais e feridos”, conforme descrito nas Tabelas 4.33 e 4.34.

Figura 4.82 – Espacialização dos acidentes “fatais e feridos” para o ano de 2015 na rodovia SP-065 do km 125 ao km 145,500.

As Figuras 4.83 e 4.84 apresentam, respectivamente, a espacialização dos acidentes “totais” e “fatais e feridos” para o ano de 2016 no trecho compreendido entre o km 125 ao km 145,500 da SP-065.

Por meio da Figura 4.83 (b) e 4.84 se verifica que as abordagens de predição utilizadas na pesquisa detectaram mais pontos críticos do que de fato foram observados, Figura 4.105 (a). Em geral, isso ocorre quando os dados não se ajustaram adequadamente ao modelo, conforme ilustrado na Figura 4.65 (a) e 4.65 (b). Sendo assim, espacialmente o modelo do HSM se mostrou mais próximo ao mapeamento das ocorrências de 2016.

Por sua vez, ao se considerar os acidentes “fatais e preditos” para o ano de 2016 se verifica que as abordagens de predição por redes exploradas na pesquisa apresentaram um bom ajuste, Figuras 4.65 (c) e 4.65 (d), e espacialização das ocorrências com este

nível de severidade, conforme ilustram as Figuras 4.84 (b) e 4.84 (c), em relação as ocorrências observadas neste ano, Figura 4.84 (a). Neste caso, as modelagens de predição propostas se mostraram mais adequadas a espacialização dos dados quando comparada ao mapeamento resultante das preditores obtidas com o modelo HSM calibrado pelo método EB, conforme ilustra a Figura 4.84 (d).

Figura 4.83 – Espacialização dos acidentes “totais” para o ano de 2016 na rodovia SP-065 do km 125 ao km 145,500.

Figura 4.84 – Espacialização dos acidentes “fatais e feridos” para o ano de 2016 na rodovia SP-065 do km 125 ao km 145,500.

Ao se analisar a variação da acidentalidade viária entre os anos de 2015 e 2016 se verifica que houve a redução do número de acidentes “totais”, e “fatais e feridos”, conforme descrito nas Tabelas 4.33 e 4.34.

4.8.1 Discussão dos resultados obtidos pelas abordagens de predição de acidentes

As espacializações obtidas por intermédio dos acidentes preditos utilizando a abordagem de predição de redes complexas bipartidas, em geral, apresentou a superestimação de pontos críticos quando comparada as espacializações dos acidentes “reais” ou observados no trecho em estudo. Verifica-se que os pontos críticos tanto para os acidentes “totais” quanto para os acidentes “fatais e feridos” começaram a ser detectados com alguns metros de antecedência, o que sugere que os acidentes nestes casos

tenham sido causados por fatores detectados nas proximidades da geolocalização do acidente fornecida pela concessionária.

Os resultados obtidos apontam que a modelagem de predição “redes simplificada” se aproxima da modelagem HSM calibrada pelo método EB, uma vez que, neste caso, ambas utilizam as mesmas variáveis preditoras, enquanto que na modelagem por “redes geral” são utilizadas as variáveis de maior peso estatístico na base de dados geral, conforme descrito na Tabela 4.6.

Nota-se que ao se acrescentar variáveis ambientais ao processo de predição, embora tenha se observado uma precisão de predição satisfatória, não foi observada, neste caso, a acurácia e a precisão dos parâmetros modelados, ou seja, um bom ajuste da base de dados ao modelo. Por sua vez, ao se considerar as variáveis recomendadas pelo manual da AASHTO (2010), que são fundamentalmente variáveis de tráfego e de infraestrutura, observa-se tanto uma melhora da precisão de predição quanto da precisão de exatidão do modelo.

Quanto melhor o ajuste dos dados ao modelo mais realista será a espacialização dos acidentes preditos e a detecção dos pontos críticos em diferentes segmentos da rodovia. Para o ano de 2010, por exemplo, houve um bom ajuste dos dados às modelagens de predição por “redes totais” e “redes simplificada”, tanto na predição dos acidentes “totais” quanto para os acidentes “fatais e feridos”, o que resultou em uma espacialização compatível à observada para este ano. Por sua vez, quanto pior o ajuste dos dados ao modelo mais distante a espacialização dos acidentes preditos estará da observada.

As abordagens de predição em redes exploradas nesta pesquisa embora apresentem o viés da precisão de exatidão quando comparada aos acidentes “reais” e ao modelo HSM calibrado pelo método EB, permitiram identificar que há uma evidente correlação espacial e temporal entre os acidentes registrados em épocas distintas.

4.9 Considerações finais

Neste capítulo foram apresentados os experimentos e os resultados lançados com base na metodologia proposta nesta pesquisa, por meio de técnicas de AM e fundamentados em conceitos de IA.

Para tanto, inicialmente realizou-se a redução da dimensionalidade da base de dados com a técnica de PCA. Desta forma, foi possível reduzir o número de variáveis utilizadas na modelagem, bem como identificar, ainda que de forma exploratória, as variáveis de maior peso estatístico, considerando o conjunto de dados em análise.

A partir das variáveis selecionadas pela técnica de PCA foram exploradas diferentes técnicas de agrupamento derivadas de conhecimentos de clustering e redes complexas homogêneas, visando reduzir a heterogeneidade da base de dados. A performance das técnicas de agrupamento foi avaliada por meio da extração de regras de decisão utilizando estruturas de dados em “árvores” (CART).

Os resultados indicaram que os agrupamentos obtidos utilizando estruturas em redes possibilitaram a detecção das estruturas mais robustas de agrupamento e a extração de regras de decisão com um maior número de variáveis associadas, com maior probabilidade de ocorrência e acerto percentual. Desta forma, foram selecionadas as variáveis que mais contribuíram à extração das regras de decisão, o que permitiu gerar uma base de dados contendo somente as variáveis essenciais a estudos de classificação e predição de acidentes rodoviários, com a finalidade de minimizar o efeito da multidimensionalidade da base de dados.

Nesta etapa, foram explorados também métodos de balanceamento de dados com a finalidade de minimizar o efeito do desbalanceamento de classes, característica inerente as bases de acidentes rodoviários, em que as ocorrências envolvendo vítimas fatais são raras quando comparadas as ocorrências de acidentes com outros níveis de severidade. Os resultados revelam que a técnica SMOTE possibilitou a amostragem dos dados de forma eficiente, sem perda de informação e contribuiu para a melhor performance dos algoritmos de classificação.

A partir dos agrupamentos de acidentes obtidos pelo algoritmo de redes complexas, as bases de dados foram classificadas usando os classificadores baseados em ANN e BNN. Os resultados indicaram que as ANN tendem a apresentar *overfitting* mesmo utilizando conjuntos de dados balanceados, ao passo que as BNN, por considerar o peso das variáveis utilizadas na modelagem, apresentam um maior potencial à detecção de falsos positivos, conforme se verifica nas matrizes de confusão.

As BNN permitiram gerar regras de classificação que contribuíram a análise microscópica de efeitos causais relacionados as variáveis de infraestrutura viária, tráfego e condições ambientais em cada segmento da rodovia em estudo.

Por fim, são apresentados os resultados obtidos utilizando diferentes abordagens de predição de acidentes rodoviários. Nota-se, que as abordagens de predição utilizando redes complexas bipartidas são mais sensíveis a dimensionalidade da base de dados, quando comparada a abordagem de predição tradicional do HSM calibrado pelo método EB.

No entanto, a precisão de predição utilizando a abordagem de redes complexas bipartidas foi superior à do HSM. Todavia, conforme já mencionado, os modelos determinísticos apresentam um maior poder discriminatório, ou seja, uma melhor precisão de exatidão ou poder de explicação, como pode-se observar na espacialização das predições.

Ressalta-se, contudo, que tanto as abordagens de predição derivadas de IA quanto a modelagem do HSM (AASHTO, 2010) apresentam falsos positivos e falsos negativos. As abordagens em redes apresentaram assim como a do HSM superestimação de acidentes. Embora a estimativa de mais acidentes em um marco quilométrico deva ser evitada, em segurança viária, modelos de previsão podem emitir diferentes níveis de alerta a concessionária, enquanto que modelos que subestimam o número de acidentes superestimam o desempenho de segurança de uma rodovia, o que pode acarretar no aumento da quantidade e severidade dos acidentes.

5 CONCLUSÕES

5.1 Síntese do trabalho

A aquisição massiva de informações vem incentivando os pesquisadores a buscarem diferentes formas de analisar problemas do mundo real, em especial, em investigações de natureza complexa nas quais os dados variam no tempo e no espaço, influenciados por eventos recorrentes, ocasionados por fontes aleatórias, como em ambiente rodoviário.

A análise e previsão de acidentes em rodovias é uma tarefa desafiadora aos pesquisadores em engenharia de transportes. Embora tenha se presenciado um crescente avanço tecnológico nas últimas décadas, não há ainda no Brasil iniciativas que permitam o amplo acesso a dados necessários a estudos desta natureza. As pesquisas desenvolvidas nesta área ficam limitadas a dados que quando fornecidos por concessionárias ou agências do governo, não apresentam as variáveis necessárias a este tipo de estudo. Em geral, os projetos são disponibilizados sem padronização ou grau de detalhamento, o que dificulta a validação de resultados de pesquisa.

Na literatura, são identificados modelos de previsão de acidentes que buscam prever, com níveis de risco satisfatórios, frequências de acidentes em pontos ou segmentos homogêneos de uma rodovia. Estes modelos visam prevenir ou reduzir o número de acidentes de tráfego, bem como o grau da severidade das lesões das vítimas no trânsito.

Em termos práticos, os modelos de previsão de acidentes rodoviários permitem que a concessionária execute intervenções na rodovia como, por exemplo, a implantação de dispositivos de controle de velocidade, a melhoria da infraestrutura viária, a fiscalização de leis e normas de trânsito.

Este trabalho de doutorado objetivou explorar algumas possibilidades em IA que permitissem detectar padrões e prever acidentes em rodovias, com precisão compatível aos métodos estatísticos tradicionais como o HSM da AASHTO (2010). Para tanto, foi proposto um protocolo de coleta de dados em ambiente rodoviário, bem como a modelagem dos dados em estruturas de clustering, árvores e redes, utilizando técnicas de agrupamento, classificação e predição de links.

O mapeamento e a revisão sistemática da literatura revelaram que as abordagens baseadas em redes são as mais utilizadas na detecção de padrões de acidentes de trânsito. Quando comparadas aos métodos estatísticos tradicionais e às modelagens em árvores (CART), elas possibilitam tratar adequadamente a multicolinearidade dos dados, fenômenos multicausais, aleatórios e não lineares.

Há na literatura uma vasta exploração de algoritmos baseados em aprendizagem supervisionada em redes, como a abordagem de ANNs e BNs. Embora sejam satisfatoriamente eficientes na classificação de padrões de severidade, as mesmas são dependentes do rótulo das classes de dados e do balanceamento da base de dados.

Desta forma, alternativamente, tem-se explorado a utilização conjunta de técnicas fundamentadas em aprendizado não supervisionado, supervisionado e semissupervisionado, com a finalidade de modelar dados rotulados e não rotulados. Ademais, tem-se considerado, quando necessário, técnicas de balanceamento de classes de dados, como a SMOTE, que possibilita a amostragem de dados, gerando amostradas sintéticas com base na distribuição original dos dados, sem perda de informação, visando estudos de classificação de risco e mensuração de níveis de severidade de acidentes de trânsito.

A aplicação prévia de técnicas de agrupamento como clustering e detecção de comunidades em redes complexas permitem reduzir a heterogeneidade dos dados e otimizar o desempenho dos classificadores por ANNs e BNs. Os resultados alcançados nesta pesquisa revelaram que a abordagem de redes complexas é mais generalista em relação a abordagem de clustering, no que se refere a detecção de estruturas de agrupamentos em bases georreferenciadas de acidentes rodoviários, apresentando uma performance de desempenho superior ao se considerar banco de dados de acidentes constituídos por diferentes níveis de severidade.

Em relação ao problema de classificação verificou-se que embora apresentem um tempo de processamento semelhante, as ANNs e as BNs exibem diferenças de modelagens substanciais. Por atribuir um peso a priori às conexões da rede, a abordagem de BNs tende a apresentar resultados mais realistas quando comparadas as ANNs, em especial, no que se refere a minimização do *overfitting* verificado na construção das matrizes de confusão.

Sendo assim, com a abordagem de BNs a partir de dados amostrados pela técnica SMOTE, é possível obter regras de classificação que consideram um maior número de variáveis e que apresentam uma maior probabilidade de ocorrência. As potencialidades das técnicas de IA diferem-se quanto a natureza do estudo, quanto ao tipo e quantidade de dados.

Ressalta-se, contudo, que os classificadores utilizados pelas ANNs e BNs realizam a previsão de rótulos sem considerar a vizinhança das observações, o que dificulta modelar padrões temporais e espaciais de acidentes, como também prever pontos críticos e estimar

níveis de desempenho de segurança numa rodovia. Essa problemática conduz a necessidade de se explorar abordagens mais sofisticadas e que ainda não tenham sido exploradas exhaustivamente na literatura no que se refere a previsão de acidentes em redes.

Neste contexto, surge ainda que emergente, na área de engenharia de transportes, a aplicação de redes complexas na predição de comunidades de acidentes e previsão de alertas de congestionamentos oriundos da ocorrência de acidentes em rodovias. No entanto, os trabalhos publicados até o presente momento exploraram redes homogêneas, ou seja, formadas por um único tipo de vértice.

Contudo, os problemas do “mundo real” são construídos por tipos de vértices distintos, como é o caso das redes de transportes. Sendo assim, nesta pesquisa, foi explorada na predição de acidentes a abordagem de redes heterogêneas bipartidas, nas quais um conjunto de vértice representa os tipos de acidentes e o outro conjunto as localizações georreferenciadas destas ocorrências ao longo do trecho de uma rodovia com características essencialmente urbanas.

A predição de acidentes utilizando predição de links pela abordagem de redes complexas bipartidas mostrou-se flexível ao número de variáveis a serem adotadas no processo de predição, prevalecendo neste caso a precisão de predição em detrimento da precisão de exatidão, como é comum na estatística tradicional que se ocupa essencialmente com o grau de explicação dos modelos. Neste trabalho, objetivou-se prever acidentes com base em correlações espaciais e temporais considerando a disponibilidade do conjunto de dados e das variáveis mensuráveis pelas concessionárias ou em *in loco*.

Com a finalidade de validar estatisticamente os resultados obtidos pela modelagem de predição de redes bipartidas, os resultados foram comparados com os valores preditos pelo modelo HSM calibrado pelo EB da AASHTO (2010). Nesta perspectiva, foram realizadas duas modelagens em redes distintas. A primeira denominada de “redes geral”, que utiliza as variáveis de maior peso da base de dados, e a segunda a modelagem denominada de “redes simplificada”, realizada a partir das variáveis recomendadas pelo manual do HSM.

Verificou-se que a modelagem por redes heterogêneas bipartidas pela modelagem de “redes simplificada” apresenta o melhor ajuste dos dados quando comparada a modelagem “redes geral”. Destaca-se que a modelagem “redes simplificada” se aproxima da HSM por utilizarem as mesmas variáveis, sendo que estas compreendem essencialmente variáveis de fluxo, infraestrutura viária e registros observacionais de

acidentes, enquanto que a modelagem “redes geral” utiliza variáveis mais complexas como condição do tempo, condição de visibilidade, inclinação da pista, etc. Embora seja mais realista considerar todas as variáveis envolvidas na ocorrência dos acidentes, se verifica uma deterioração da qualidade da precisão da modelagem quando comparada a modelagem de “redes simplificada” para as bases de dados analisadas.

A partir da seleção das variáveis e técnicas necessárias à realização deste estudo, foi possível realizar uma análise comparativa das diferentes técnicas de IA, com base em cada domínio de aprendizagem, bem como avaliar sua precisão e acurácia na investigação e predição de acidentes.

Por compreender um modelo paramétrico, a modelagem do HSM é menos afetada pela dimensionalidade da base de dados, quando comparada a modelagem em redes. No entanto, ressalta-se que o trecho em análise ainda se encontra em obras, o que influencia diretamente na detecção de padrões de acidentes pelas técnicas de IA.

Embora limitada pela disponibilidade dos dados fornecidos pela concessionária para a realização desta pesquisa, verifica-se que a mesma apresenta resultados promissores. Foi possível identificar correlações espaciais e temporais, bem como os pontos críticos no trecho em estudo. Todas as abordagens de predição utilizadas apresentaram superestimação do número de acidentes, ou seja, falso positivos, inclusive a HSM para a base de dados analisada.

Contudo, embora devam ser evitados os “falsos positivos”, podem servir de alertas para as concessionárias e agências fiscalizadoras, no que se refere a identificação de áreas com alta densidade de acidentes. Cada acidente previsto adicionalmente é um acidente que pode ser evitado, o que não é possível em modelagens que subestimam o número de acidentes e, conseqüentemente, superestimam os níveis de desempenho e segurança de uma rodovia.

5.2 Considerações finais e contribuições

Foi possível por meio das modelagens de investigação e predição propostas neste trabalho prever, com níveis satisfatórios de acurácia e precisão, a frequência de acidentes em rodovias. Quanto aos objetivos principais da tese, concluiu-se que por agregar conceitos de análise de vizinhança, a modelagem proposta fornece informações essenciais a detecção de padrões, predição e prevenção de acidentes futuros. Logo, abordagens que explorem estes conceitos devem ser consideradas em estudos de investigação da acidentalidade viária.

A modelagem considerando a abordagem de redes bipartidas georreferenciadas se mostra flexível em relação ao número de variáveis utilizadas no processo de modelagem. Foi possível compreender a variação entre os vários fatores envolvidos na ocorrência dos acidentes (motorista, veículo, via e meio ambiente), enquanto que as abordagens estatísticas tradicionais compreendem soluções fechadas, em que pressupostos devem ser estabelecidos a priori entre as variáveis utilizadas na modelagem.

Ressalta-se que os resultados alcançados por meio da investigação e predição de acidentes rodoviários utilizando técnicas de IA, trouxeram contribuições à área de engenharia de transportes, dentre as quais: (i) Proposição de um protocolo de coleta de dados em ambiente rodoviário; (ii) Detecção de padrões e severidade de acidentes de tráfego por meio de técnicas de agrupamentos baseados em clustering e redes complexas, utilizando indicadores de desempenho como a silhueta e a modularidade, respectivamente, bem como o somatório dos erros quadráticos residuais; (iii) Análise de diferentes técnicas de amostragem de bases de dados desbalanceadas como a *undersampling*, *oversampling* e SMOTE; (iv) Avaliação da performance dos algoritmos de agrupamento por meio da extração de regras de decisão utilizando a modelagem de dados em árvores binárias (CART); (v) Obtenção de regras de classificação pela modelagem de dados em estruturas de redes, pela abordagem de ANNs e BNs; (vi) Predição de acidentes em rodovias a partir da modelagem de dados em redes complexas heterogêneas bipartidas, considerando conceitos baseados em vizinhança e predição de links em redes; (vii) Validação dos resultados obtidos por meio do ajustamento das observações preditas utilizando o teste estatístico *cure plot*; (viii) Comparação dos resultados obtidos com modelo HSM calibrado pelo EB, de acordo com o manual da AASHTO (2010).

Contudo, como mencionado anteriormente, a pesquisa ficou limitada a quantidade de dados fornecidos pela concessionária. Sendo que os dados do período de depois, apresentam ruídos ou outliers, como verificado na etapa de ajustamento das observações. Esta particularidade dificulta a detecção de padrões e a predição de acidentes, uma vez que o trecho da rodovia em estudo ainda se encontra em obras.

5.3 Recomendações

Como recomendação para trabalhos futuros sugere-se que sejam analisados dados com uma variação espacial e temporal superior a utilizada na pesquisa, bem como diferentes

tipos de rodovias (urbana ou rural), considerando alterações no número de faixas, na quantidade de dispositivos de acesso e na densidade de aglomerados urbanos.

Como alternativa a restrita disponibilidade de dados para a realização de pesquisas desta natureza, pode-se explorar dados oriundos do BIM (*Building Information Model*) de simuladores de tráfego ou de imagens biométricas, que proporcionam a modelagem da informação presente nos projetos de engenharia referentes as rodovias analisadas. Estas ferramentas permitem simular “*hot spots*”, ou seja, pontos de perigo na rodovia, considerando variações espaço-temporais, por meio de técnicas de IA fundamentadas em aprendizado supervisionado e semissupervisionado.

Deve-se considerar também as variações geomorfológicas da região em que se encontra inserida uma rodovia. Desta forma, variáveis associadas ao clima, ou seja, ao meio ambiente, podem ser modeladas e especificidades regionais ou locais podem ser detectadas.

Além disso, uma variável pode ser mais ou menos expressiva para diferentes padrões de rodovia, sendo assim essa particularidade deve ser considerada no processo de modelagem dos dados. Como, por exemplo, tem-se a variável iluminação artificial, que em rodovias urbanas matematicamente não fornece impacto na modelagem dos dados, uma vez que nestes casos a iluminação da rodovia é de responsabilidade do município circunjacente a rodovia, conforme estabelecido em plano diretor, o que implica que a contagem de desta variável implantada pela concessionária ao longo da rodovia não seja tão expressiva.

Outro aspecto importante refere-se ao formato, tipo e a quantidade de dados. Desta forma, recomenda-se testar diferentes algoritmos de agrupamentos e classificadores multiclases baseados em redes, bem como diferentes medidas de similaridades e critérios baseados em vizinhança.

Ademais, é relevante que sejam explorados outros coeficientes de desempenho que permitam avaliar a performance dos algoritmos de detecção de comunidades, classificação e previsão de acidentes utilizando a abordagem em redes. O estabelecimento de marcadores de desempenho pode auxiliar na análise de grandes e heterogêneas bases de dados.

Por fim, é necessário que se sejam criados conjuntos de dados de referência, denominados de *benchmarking*. Estes conjuntos de dados permitem testar com margem de segurança diferentes técnicas de IA evitando-se assim padrões anômalos nos dados, como foi observado nesta pesquisa.

REFERÊNCIAS

- AASHTO. **Highway Safety Manual**, Washington: American Association of State and Highway Transportation Officials, First Edition, EUA, 2010.
- Abellán, J.; López, G.; De Oña, J. Analysis of traffic accident severity using Decision Rules via Decision Trees. **Expert Systems with Applications**, v. 40, n. 15, p. 6047–6054, 2013.
- Abdi, H.; Williams, L. J. Principal component analysis. **Wiley Interdisciplinary Reviews: Computational Statistics**, Volume 2, 2010.
- Adamic, L. A.; Adar, E. Friends and neighbour's on the Web. **Social Networks**, v. 25, n. 3, p. 211–230, 2003.
- Ahmed, M.; Huang, H.; Abdel-Aty, M.; Guevara, B. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. **Accident Analysis and Prevention**, v. 43, n. 4, p. 1581–1589, 2011.
- Alejo, R.; Valdovinos, R. M.; García, V.; Pacheco-Sanchez, J. H. A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. **Pattern Recognition Letters**, v. 34, n. 4, p. 380–388, 2013.
- Anand, R.; Mehrotra, K. G.; Mohan, C. K.; Ranka, S. An improved algorithm for neural network classification of imbalanced training sets. **IEEE transactions on neural networks**, v. 4, n. 6, p. 962–969, 1993.
- Anastasiu, D. C.; Karypis, G. L2KNNg: Fast Exact K-Nearest Neighbor Graph Construction with L2_Norm Pruning. In: **24th ACM International Conference on Information and Knowledge Management**. CIKM'15, Melbourne, Australia, 2015, p. 1-10.
- Anderson, T. K. Kernel density estimation and K-means clustering to profile road accident hotspots. **Accident Analysis and Prevention**, v. 41, n. 3, p. 359–364, 2009.
- Alejo, R.; Valdovinos, R. M.; García, V.; Pacheco-Sanchez, J. H. A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. **Pattern Recognition Letters**, v. 34, n. 4, p. 380–388, 2013.
- Bao, Z. K.; MA, C.; XIANG, B.; ZHANG, H. Identification of influential nodes in complex networks: Method from spreading probability viewpoint. **Physica A: Statistical Mechanics and its Applications**, v. 468, p. 391–397, 2017.
- Barabási, A. L.; Jeong, H.; Neda, Z.; Ravasz, E.; Schubert, A.; Vicsek, T. Evolution of the social network of scientific collaboration. **Physica A: Statistical Mechanics and its Applications**, v. 311, n. 3, p. 590–614, 2002.

- Barabási, A. L.; Albert, R. Emergence os Scaling in Random Networks. **Science**, v. 286, n. 5439, p. 509–512, 1999.
- Beckett, S. J. Improved community detection in weighted bipartite networks. **The Royal Society**, v. 3, n. 1, 2016.
- Berton, L. **Construção de redes baseadas em vizinhança para o aprendizado semissupervisionado**. Tese (Doutorado – Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional). Universidade de São Paulo: São Carlos, 2016. 135p.
- Berton, L.; Faleiros, T.; Valejo, A. Valverde-Rebaza, J.; Lopez, A. A. RGCLI: Robust Graph that Considers Labeled Instances for Semi-Supervised Learning. **Neurocomputing**, v. 226, p. 238–248, 2017.
- Bolón-Canedo, V.; Sánchez-Marono, N.; Alonso-Betanzos, A.; Benítez, J. M.; Herrera, F. A review of microarray datasets and applied feature selection methods. **Information Sciences**, v. 282, p. 111-135, 2014.
- Bonacich, P. Power and Centrality: A Family of Measures. **American Journal of Sociology**, v. 92, n. 5, p. 1170–1182, 1987.
- Braga, A. P.; Carvalho, A. C. P. L. F.; Ludermir, T. B. Redes Neurais Artificiais. In: Org. Rezende, S. O. **Sistemas inteligentes: fundamentos e aplicações**. Barueri - SP: Manole, 2005. 525p.
- Brandes, U. A faster algorithm for betweenness centrality. **The Journal of Mathematical Sociology**, v. 25, n. 2, p. 163–177, 2001.
- Caliendo, C.; Guida, M.; Parisi, A. A crash-prediction model for multilane roads. **Accident Analysis and Prevention**, v. 39, n. 4, p. 657–670, 2007.
- Castellanos, F. J.; Valero-Mas, J. J.; Calvo-Zaragoza, J.; Rico-Juan, J. R. Oversampling imbalanced data in the string space. **Pattern Recognition Letters**, v. 103, p. 32–38, 2018.
- Cateni, S.; Colla, V.; Vannucci, M. Neurocomputing A method for resampling imbalanced datasets in binary classification tasks for real-world problems. **Neurocomputing**, v. 135, p. 32–41, 2014.
- Chapelle, O.; Schlkopf, B.; Zien, A. **Semi-supervised learning**. 1st ed. The MIT Press, 2010.
- Chang, L-Y. Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. **Safety Science**, v. 43, p. 541-557, 2005.
- Chang, L.; Chen, W. Data mining of tree-based models to analyze freeway accident frequency. **Journal of Safety Research**, v. 36, p. 365–375, 2005.

- Chang, L.; Wang, H. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. **Accident Analysis and Prevention**, v. 38, p. 1019–1027, 2006.
- Chawla, Nitesh V.; Bowyer, Kevin W.; Hall, Lawrence O.; Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique, **Journal of Artificial Intelligence Research**, v. 16, p. 321–357.
- Chen, C.; Zhang, G.; Yang, J.; Milton, J. C.; Alcántara, A. D. An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. **Accident Analysis and Prevention**, v. 90, p. 95–107, 2016.
- Chen, W.; Guo, F.; Wang, F.-Y. A Survey of Traffic Data Visualization. **IEEE Transactions on Intelligent Transportation Systems**, v. 16, n. 6, p. 2970–2984, 2015.
- Chuerubim, M. L.; Valejo, A.; Silva, I. Spatial and temporal identification of community structures of road accidents through homogeneous complex networks and measures of centrality. In: **8º Congresso Luso-Brasileiro para o Planeamento Urbano, Regional, Integrado E Sustentável, Cidades e Territórios - Desenvolvimento, atratividade e novos desafios**, PLURIS 2018, pp. 1–12, Coimbra, Portugal, 2018.
- Clauset, A., Newman, M. E., Moore, C. Finding community structure in very large networks, **Physical Review E**, Vol 70, Nº. 6, 2004.
- Costa, J. O.; Jacques, M. A. P.; Soares, F. E. C.; Freitas, E. F. Integration of geometric consistency contributory factors in three-leg junctions collision prediction models of Portuguese two-lane national highways. **Accident Analysis and Prevention**, v. 86, p. 59–67, 2016.
- Costa, L. F.; Rodrigues, F. A.; Travieso, G.; Villas Boas, P. R. Characterization of Complex Networks: A Survey of measurements. **Advances in Physics**, v. 56, n. 1, p. 167–242, 2007.
- Csárdi, G.; Nepusz, T. The igraph software package for complex network research. **InterJournal Complex Systems**, v. 1695, p. 1–9, 2006.
- Cui, W.; Pu C.; Xu, Z. Bounded link prediction in very large networks. **Physica A: Statistical Mechanics and its Applications**, v. 457, p. 202–214, 2016.
- Delen, D.; Sharda, R.; Bessonov, M. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. **Accident Analysis & Prevention**, v. 38, p. 434–444, 2006.
- De Oña, J.; De Oña, R.; Eboli, L.; Forciniti, C.; Machado, J. L.; Mazzulla, G. Analysing the Relationship Among Accident Severity, Drivers' Behaviour and Their Socio-economic Characteristics in Different Territorial Contexts. **Procedia - Social and Behavioral Sciences**, v. 160, n. Cit, p. 74–83, 2014.

- De Oña, J.; López, G.; Mujalli, R.; Calvo, F. J. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. **Accident Analysis and Prevention**, v. 51, p. 1–10, 2013.
- De Oña, J.; Mujalli, R. O.; Calvo, F. J. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. **Accident Analysis and Prevention**, v. 43, n. 1, p. 402–411, 2011.
- Depaire, B.; Wets, G.; Vanhoof, K. Traffic accident segmentation by means of latent class clustering. **Accident Analysis and Prevention**, v. 40, n. 4, p. 1257–1266, 2008.
- DeTienne, K.B.; DeTienne, D.H.; Joshi, S. A. Neural Networks as Statistical Tools for Business Researchers. **Organizational Research Methods**, v. 6, p. 236–245, 2003.
- Ding, J.; Jiao, L.; Wu, J.; Liu, F. Prediction of missing links based on community relevance and ruler inference. **Knowledge-Based Systems**, v. 98, p. 200–215, 2016.
- Dorogovtsev, S. N.; Mendes, J. F. F. Evolution of Networks: From Biological Nets to the Internet and WWW. **Booksgooglecom**, v. 57, p. 280, 2003.
- Facelli, K., Lorena, A. C., Gama, J., Carvalho, A.C.P.L.F. **Inteligência Artificial: Uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2011. 378p.
- Fawcett, T. **Learning from Imbalanced Classes**. 2016. Disponível em: <<https://www.svds.com/learning-imbalanced-classes/>>. Acesso: novembro/2018.
- Fawcett, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, p. 861–874, 2006.
- Ferraz, A. C. P.; Raia Júnior, A. Bezerra, B.; Bastos, T.; Rodrigues, K. **Segurança no Trânsito**. São Carlos: Grupo Gráfico São Francisco, 2008. 280p.
- Fielding, A. H.; Bell, J. F. A review of methods for the assessment of prediction errors in conservation presence absence models. **Environmental Conservation**, v. 24, n. 1, p. 38–49, 1997.
- Fiorini, P. C.; Jabbour, C. J. C. Information systems and sustainable supply chain management towards a more sustainable society: Where we are and where we are going. **International Journal of Information Management**, v. 37, n. 4, p. 241–249, 2017.
- Freeman, L. C. **A Set of Measures of Centrality Based on Betweenness Sociometry**, 1977. Disponível em: <<http://www.jstor.org/stable/3033543?origin=crossref>>. Acesso: novembro/2018.

- Fu, Y.; Li, C.; Luan, T. H.; Zhang, Y.; Mao, G. Infrastructure-cooperative algorithm for effective intersection collision avoidance. **Transportation Research Part C**, v. 89, n. July 2017, p. 188–204, 2018.
- Galar, M.; Fernández, A.; Barrenechea, E.; Bustince, H.; Herrera, F. Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets. **Information Sciences**, v. 354, p. 178–196, 2016.
- Getoor, L.; Diehl, C. P. Link mining: A survey. **ACM SIGKDD Explorations Newsletter**, v. 7, n. 2, p. 3–12, 2005.
- Google Earth. **Imagens Quickbird da Rodovia Dom Pedro Primeiro (SP- 065)**. Google: Digital Globe, 2018.
- Guillaume, J.; Latapy, M. Bipartite Structure of All Complex Networks Bipartite Structure of all Complex Networks. **Information Processing Letters**, v. 90, p. 215–221, 2004.
- Guillaume, J.; Latapy, M. Bipartite Graphs as Models of Complex Networks. **Physica A**, v. 371, p. 795–813, 2006.
- Gupta, A.; Lam, M. S. Estimating missing values using neural networks. **Journal of the Operational Research Society**, v. 47, p. 229–238, 1996.
- Hair, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E. **Multivariate data analysis**. New York: Prentice Hall, 2009. 816p.
- Halim, Z.; Kalsoom, R.; Bashir, S.; Abbas, G. Artificial intelligence techniques for driving safety and vehicle crash prediction. **Artificial Intelligence Review**, v. 46, n. 3, p. 351–387, 2016.
- Han, J.; Kamber, M.; Pei, J. Mining heterogeneous information networks by exploring the power of links. **Lecture Notes in Computer Science**, v. LNAI 5808, n. Springer-Verlag Berlin Heidelberg, p. 13–30, 2009.
- Hauer, E. **The Art of Regression Modeling in Road Safety**. University of Toronto. Canada. 2014. pp. 1–199.
- Hauer, E.; Bamfo, J. Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables. In: **Proceedings ICTCT (International Cooperation on Theories and Concepts in Traffic Safety)**, Lund, Sweden, 1997.
- Herlocker, J. L.; Konstan, J. A.; Terveen, L. G.; Riedl, J. T. Evaluating collaborative filtering recommender systems. **ACM Transactions on Information Systems (TOIS)**, v. 22, n. 1, p. 5–53, 2004.
- Hongyu, K., Sandanielo, V. L. M.; Oliveira Júnior, G. J. Principal Component Analysis: theory, interpretations and applications. **E & S – Engineering and Science**, Volume1, 5:1 ed., 2015.

- Hosmer, D.W.; Lemeshow, S. **Applied logistic regression**, 2nd Ed. John Wiley & Sons, New York, 2000.
- Huang, H.; Abdel-Aty, M. Multilevel data and Bayesian analysis in traffic safety. **Accident Analysis and Prevention**, v. 42, n. 6, p. 1556–1565, 2010.
- Huang, J.; Charles, X. L. Using AUC and Accuracy in Evaluating Learning Algorithms. **IEEE Transactions on Knowledge and Data Engineering**, v. 17, n. 3, p. 299–310, 2005.
- Instituto Brasileiro de Geografia e Estatística (IBGE). Rede Brasileira de Monitoramento Contínuo dos Sistemas GNSS (RBMC. IBGE: Rio de Janeiro, 2018. Disponível em: https://downloads.ibge.gov.br/downloads_geociencias.htm . Acesso: julho/2018.
- Iranitalab, A.; Khattak, A. Comparison of four statistical and machine learning methods for crash severity prediction. **Accident Analysis and Prevention**, v. 108, n. February, p. 27–36, 2017.
- Jabbour, C. J. C. Environmental training in organisations: From a literature review to a framework for future research. **Resources, Conservation and Recycling**, v. 74, p. 144–155, 2013.
- Johnson, R. A.; Wichern, D. W. **Applied multivariate statistical analysis**. 6th ed. Upper Saddle River, New Jersey: Prentice-Hall, 2002.
- Karlaftis, M. G.; Vlahogianni, E. I. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. **Transportation Research Part C**, v. 19, n. 3, p. 387–399, 2011.
- Kashani, A. T.; Shariat-Mohaymany, A.; Ranjbari, A. A data mining approach to identify key factors of traffic injury severity. **Traffic & Transportation**, v. 23, p. 11–17, 2011.
- Kaufman, L.; Rousseeuw, P. J. Finding Groups in Data: An Introduction to Cluster Analysis. **Biometrics**, v. 47, n. 2, p. 788, 1991.
- Khoo, H. L.; Ahmed, M. Modeling of passengers' safety perception for buses on mountainous roads. **Accident Analysis and Prevention**, v. 113, n. January, pp. 106–116, 2018.
- Khosravi, H.; Bina, B.; A Survey on Statistical Relation Learning. In: **AI'10: Proceedings of the 23rd Canadian Conference on Advanced in Artificial Intelligence**, 2010, p. 256-268.
- Kitchenham, B.; Brereton, O. P.; Budgen, D.; Turner, M.; Bailey, J.; Linkman, S. Systematic literature reviews in software engineering - A systematic literature review. **Information and Software Technology**, v. 51, n. 1, p. 7–15, 2009.

- Kitchenham, B. **Procedures for performing systematic reviews**. Keele, UK, Keele University, 33, p. 1–26, 2004.
- Kirby, H., Dougherty, M., Watson, S. Should we use neural networks or statistical models for short term motorway forecasting? **International Journal of Forecasting**, v. 13, p. 45–50, 1997.
- Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. **Appears in the International Joint Conference on Artificial Intelligence (IJCAI)**, v. 5, p. 1–7, 1995.
- Krawczyk, B. Learning from imbalanced data: open challenges and future directions. **Progress in Artificial Intelligence**, v. 5, n. 4, p. 221–232, 2016.
- Krzanowski, W. J.; Marriott, F. H. C. **Multivariate Analysis**. Part 2. John Wiley, 1a. Ed., 1995.
- Kumar, S.; Toshniwal, D.; Parida, M. A comparative analysis of heterogeneity in road accident data using data mining techniques. **Evolving Systems**, v. 8, n. 2, p. 147–155, 2017.
- Kumar, S.; Toshniwal, D. A novel framework to analyze road accident time series data. **Journal of Big Data**, v. 3, n. 1, 2016.
- Kumar, S.; Toshniwal, D. A data mining framework to analyze road accident data. **Journal of Big Data**, v. 2, n. 1, 2015.
- Kunegis, J. Exploiting The Structure of Bipartite Graphs for Algebraic and Spectral Graph Theory Applications. **Internet Mathematics**, v. 11, n. 3, p. 201–321, 2014.
- Kwon, O. H.; Rhee, W.; Yoon, Y. Application of classification algorithms for analysis of road safety risk factor dependencies. **Accident Analysis and Prevention**, v. 75, p. 1–15, 2015.
- Lawrence, S.; Burns, I.; Back, A. D. C.; Tsoi, A. C.; Lee, G. Neural Network Classification and Prior Class Probabilities, State-of-the-Art Surveys. **Tricks of the Trade, Lect. Notes in Computer. Sci.**, n. 299, p. 299–314, 1998.
- LeSage, J., Pace, R. **Introduction to Spatial Econometrics**. Chapman & Hall/CRC, Taylor & Francis Group, 2009.
- Li, J.; Fong, S.; Wong, R. K.; Chu, V. W. Adaptive multi-objective swarm fusion for imbalanced data classification. **Information Fusion**, v. 39, p. 1–24, 2018.
- Li, J.; Fong, S.; Wong, R. K.; Mohammed, S.; Fiaidhi, J.; Sung, Y. A suite of swarm dynamic multi-objective algorithms for rebalancing extremely imbalanced datasets. **Applied Soft Computing Journal**, p. 1–22, 2017.

- Li, X.; Lord, D.; Zhang, Y.; Xie, Y. Predicting motor vehicle crashes using Support Vector Machine models. **Accident Analysis and Prevention**, v. 40, n. 4, p. 1611–1618, 2008.
- Lin, L.; Wang, Q.; Sadek, A. Data Mining and Complex Networks Algorithms for Traffic Accident Analysis. **Transportation Research Record: Journal of the Transportation Research Board**, TRB Paper, v. 2460, N. 14-4172, p. 128-136, 2014.
- Li, Z.; Wang, W.; Liu, P.; Bigham, J. M.; Ragland, D. R. Using Geographically Weighted Poisson Regression for county-level crash modeling in California. **Safety Science**, v. 58, p. 89–97, 2013.
- Liu, J.; Khattak, A. J.; Wali, B. Do safety performance functions used for predicting crash frequency vary across space? Applying geographically weighted regressions to account for spatial heterogeneity. **Accident Analysis and Prevention**, v. 109, n. October, p. 132–142, 2017.
- Lobo, J. M.; Jiménez-Valverde, A.; Real, R. AUC: A misleading measure of the performance of predictive distribution models. **Global Ecology and Biogeography**, v. 17, n. 2, p. 145–151, 2008.
- López, G.; De Oña, J.; Garach, L.; Baena, L. Influence of deficiencies in traffic control devices in crashes on two-lane rural roads. **Accident Analysis and Prevention**, v. 96, p. 130–139, 2016.
- Loyola-González, O.; Martínez-Trinidad, J. F.; Carrasco-Ochoa, J. A.; García-Borroto, M. Neurocomputing Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. **Neurocomputing**, v. 175, p. 935–947, 2016.
- Lorrain, F.; White, H. Structural Equivalence of Individuals in Social Networks. **Journal of Mathematical Sociology**, v. 1, n. 1, p. 49–80, 1971.
- Lü, L.; Zhou, T. Link prediction in weighted networks: The role of weak ties. **EPL (Europhysics Letters)**, v. 89, n. 1, p. 18001, 2010.
- Macskassy, S. A.; Provost, F. Classification in Networked Data A Toolkit and a Univariate Case Study. **Journal of Machine Learning Research**, v. 8, p. 935–983, 2007.
- Maloof, M. A. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. **Analysis**, v. 21, n. Ii, p. 1263–1284, 2003.
- Mannering, F. L.; Bhat, C. R. Analytic methods in accident research: Methodological frontier and future directions. **Analytic Methods in Accident Research**, v. 1, p. 1–22, 2014.
- Mardia, K. V.; Kent, J. T.; Bibby, J. M. **Multivariate Analysis**. New York: Academic Press, 1979.

- Marhavilas, P. K.; Koulouriotis, D.; Gemeni, V. Risk analysis and assessment methodologies in the work sites: On a review, classification and comparative study of the scientific literature of the period 2000-2009. **Journal of Loss Prevention in the Process Industries**, v. 24, n. 5, p. 477–523, 2011.
- Mitchell, T. M. **Machine Learning**. McGraw-Hill Education, 1997.
- Moghaddam, A. M.; Tabibi, Z.; Sadeghi, A.; Ayati, E.; Ravandi, A. G. Screening out accident-prone Iranian drivers: Are their at-fault accidents related to driving behavior? **Transportation Research Part F: Traffic Psychology and Behaviour**, v. 46, p. 451–461, 2017.
- Monard, M. C.; Baranauskas, J. A. Conceitos sobre Aprendizado de Máquina. In: Org. Rezende, S. O. **Sistemas inteligentes: fundamentos e aplicações**. Barueri - SP: Manole, 2005. 525p.
- Mujalli, R. O.; López, G.; Garach, L. Bayes classifiers for imbalanced traffic accidents datasets. **Accident Analysis and Prevention**, v. 88, p. 37–51, 2016.
- Mussone, L.; Ferrari, A.; Oneta, M. An analysis of urban collisions using an artificial intelligence model. **Accident Analysis and Prevention** 31, v. 31, p. 705–718, 1999.
- Newman, M. E. J. **Networks: An Introduction**. 1. ed. New York: Oxford University Press, 2010.
- Newman, M. E. J. Coauthorship networks and patterns of scientific collaboration. In: **Proceedings of the National Academy of Science of the United States**, v. 101, n. PNAS'04, p. 5200–5205, 2004a.
- Newman, M. E. J. Analysis of weighted networks. **Physical Review E - Statistical, Nonlinear, and Soft Matter Physics**, v. 70, n. 5 2, p. 1–9, 2004b.
- Newman, M. E. J. Fast Algorithm for Detecting Community Structure in Networks Networks. **Physical Physical Review E**, v. 69, p. 1-5, 2004c.
- Newman, M. E. J. The Structure and Function of Complex Networks. **SIAM Review**, v. 45, n. 2, p. 167–256, 2003.
- Newman, M. E. J. The structure of scientific collaboration networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98, n. 2, p. 404–409, 2001.
- Nitsche, P.; Thomas, P.; Stuetz, R.; Welsh, R. Pre-crash scenarios at road junctions: A clustering method for car crash data. **Accident Analysis and Prevention**, v. 107, n. July, p. 137–151, 2017.

- Nowakowska, M. Selected aspects of prior and likelihood information for a Bayesian classifier in a road safety analysis. **Accident Analysis and Prevention**, v. 101, p. 97–106, 2017.
- Pande, A.; Abdel-Aty, M. Assessment of freeway traffic parameters leading to lane-change related collisions. **Accident Analysis & Prevention**, v. 38, p. 936–948, 2006.
- Papadimitriou, A.; Symeonidis, P.; Manolopoulos, Y. Fast and accurate link prediction in social networking systems. **Journal of Systems and Software**, v. 85, n. 9, p. 2119–2132, 2012.
- Park, J.; Abdel-Aty, M. Development of adjustment functions to assess combined safety effects of multiple treatments on rural two-lane roadways. **Accident Analysis and Prevention**, v. 75, p. 310–319, 2015.
- Peng, W. et al. Link Prediction in Social Networks: the State-of-the-Art. **Sci China Inf Sci**, v. 58, n. 58, p. 11101–38, 2015.
- Permai, S. D.; Ulama, B. S. S.; Iriawan, N. **Bayesian Approach for Neural Network**, v. 53, n. 4, 2015.
- Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic mapping studies in software engineering. In: **12th International Conference on Evaluation and Assessment in Software Engineering**, Vol. 17, No. 1, 2008.
- Plug, C.; Xia, J.; Caulfield, C. Spatial and temporal visualisation techniques for crash analysis. **Accident Analysis and Prevention**, v. 43, n. 6, p. 1937–1946, 2011.
- Piri, s.; Delen, D.; Liu, T. A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. **Decision Support Systems**, v. 106, p. 15–29, 2018.
- Principe, J. C., Euliano, N. R., Lefebvre, C.W. **Neural and Adaptive Systems: Fundamentals Through Simulations**. John Wiley and Sons Inc, 2000.
- QGIS. **Geographic Information System. Open Source**. Versão 2.14.18. Disponível em: <<https://qgis.org/en/site/>>. Acesso: novembro/2018.
- Rezende, S. O.; Pugliesi, J. B.; Melanda, E. A.; Paula, M. F. Mineração de dados. In: Org. Rezende, S. O. **Sistemas inteligentes: fundamentos e aplicações**. Barueri - SP: Manole, 2005. 525p.
- Robins, G.; Alexander, M. Small Worlds Among Interlocking Directors: Network Structure and Distance in Bipartite Graphs, **Computational and Mathematical Organization Theory**, 2004, v. 10, n. 1, p. 69-94.
- Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, 53–65, 1987.

- Rota das Bandeiras. **Banco de dados de acidentes da rodovia Dom Pedro I (SP -065) no período de 2009 a 2016**. Rota das Bandeiras: Grupo Odebrecht, 2016.
- Roy, A.; Cruz, R. M. O.; Sabourin, R.; Cavalcanti, G. D. C. Neurocomputing A study on combining dynamic selection and data preprocessing for imbalance learning. **Neurocomputing**, v. 286, p. 179–192, 2018.
- Sabidussi, G. The centrality index of a graph. **Psychometrika**, v. 31, n. 4, p. 581–603, 1966.
- Shadroo, S.; Rahmani, A. M. Systematic survey of big data and data mining in internet of things. **Computer Networks**, v. 139, p. 19–47, 2018.
- Salunkhe, U. R.; Mali, S. N. Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach. **Procedia Computer Science**, v. 85, n. Cms, p. 725–732, 2016.
- Saha, D.; Alluri, P.; Gan, A. Prioritizing Highway Safety Manual’s crash prediction variables using boosted regression trees. **Accident Analysis and Prevention**, v. 79, p. 133–144, 2015.
- Saha, D.; Alluri, P.; Gan, A. A random forests approach to prioritize Highway Safety Manual (HSM) variables for data collection. **Journal of Advanced Transportation**, v. 47, n. June 2010, p. 512–525, 2011.
- Sain, H.; Wulan, S. Combine Sampling Support Vector Machine for Imbalanced Data Classification. **Procedia Computer Science**, v. 72, p. 59–66, 2015.
- Soares, P. R. DA S.; Prudêncio, R. B. C. Time series based link prediction. WCCI 2012 **IEEE World Congress on Computational Intelligence**, Brisbane, Australia, June, p. 10-15, 2012,
- Sohn, S. Y.; Lee, S. H. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. **Safety Science**, v. 41, n. 1, p. 1–14, 2003.
- Solanas, A.; Manolov, R.; Leiva, D. Retaining principal components for discrete variables. **Anuario de Psicologia, Facult de Psicologia**, Universitat de Barcelona, vol. 41, n. 1-3, p. 33-50, 2011.
- Spanurattana, S.; Murata, T. Visual Analysis of Bipartite Networks. In: **2011 IEEE 11th International Conference on Data Mining Workshops**, 2011, p. 833-838.
- Sun, L.; Liu, L.; Xu, Z.; Jie, Y.; Wei, D.; Wang P. Locating inefficient links in a large-scale transportation network. **Physica A: Statistical Mechanics and its Applications**, v. 419, p. 537–545, 2015.

- Teodorović, D. Swarm intelligence systems for transportation engineering: Principles and applications. **Transportation Research Part C: Emerging Technologies**, v. 16, n. 6, p. 651–667, 2008.
- Theofilatos, A.; Yannis, G. A review of the effect of traffic and weather characteristics on road safety. **Accident Analysis and Prevention**, v. 72, p. 244–256, 2014.
- Thill, J. C. Geographic information systems for transportation in perspective. **Transportation Research Part C: Emerging Technologies**, v. 8, n. 1, p. 3–12, 2000.
- Ting, Z.; Bin, W.; Qi, G. Congestion warning method based on the Internet of vehicles and community discovery of complex networks. **Journal of China Universities of Posts and Telecommunications**, v. 23, pp. 37–45, 2016.
- Valejo, A.; Ferreira, V.; Geraldo, P. R.; Filho, G. P. R.; Oliveira, M. C. F.; Lopes, A. A. One-mode projection-based multilevel approach for community detection in bipartite networks. In: **SIMBig Track on Social Network and Media Analysis and Mining (SNMAM)**, v. 2029, p. 101–108, 2017a.
- Valejo, A. D.; Ferreira, M. C. O.; Filho, G. P. R.; Lopes, A. A. Multilevel approach for combinatorial optimization in bipartite networks. **Knowledge-Based Systems**, v. 151, p. 45–61, 2018.
- Venkataraman, N. S.; Ulfarsson, G. F.; Shankar, V. N. Extending the Highway Safety Manual (HSM) framework for traffic safety performance evaluation. **Safety Science**, v. 64, p. 146–154, 2014.
- Wach, W. Calculation reliability in vehicle accident reconstruction. **Forensic Science International**, v. 263, pp. 27–38, 2016.
- Wang, X.; Qin, X. Asymmetric intimacy and algorithm for detecting communities in bipartite networks. **Physica A**, v. 462, p. 569–578, 2016.
- Warner, B.; Misra, M., 1996. Understanding neural networks as statistical tools. **The American Statistician** 50, v. 4, p. 284–293, 1996.
- Wasserman, S.; Faust, K. **Social network analysis: methods and applications**. Cambridge University Press, 1994. 825p.
- Wei, X.; Shu, X.; Huang, B.; Taylor, E. L.; Chen, H. Analyzing Traffic Crash Severity in Work Zones under Different Light Conditions. **Journal of Advanced Transportation**, v. 2017, 2017.
- Weiss, G. M. Mining with Rarity: A Unifying Framework. **SIGKDD Explorations**, v. 6, n. 1, p. 7–19, 2004.
- Xie, Y.; Lord, D.; Zhang, Y. Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. **Accident Analysis and Prevention**, v. 39, n. 5, p. 922–933, 2007.

- Xu, Y.; Chen, L.; Li, B. Density-based modularity for evaluating community structure in bipartite networks. **Information Sciences**, v. 317, p. 278–294, 2015.
- Yang, Q.; Wu, X. 10 challenging problems in data mining research. **Internat. J. Inform. Technol. Decis. Making**, v. 5, n. 4, p. 597–604, 2006.
- Yao, S.; Loo, B. P. Y.; Yang, B. Z. Traffic collisions in space: four decades of advancement in applied GIS. **Annals of GIS**, v. 22, n. 1, p. 1–14, 2016.
- Young, J.; Park, P. Y. Benefits of small municipalities using jurisdiction-specific safety performance functions rather than the Highway safety Manual's calibrated or uncalibrated safety performance functions. **Canadian Journal of Civil Engineering**, v. 40, n. 6, p. 517–527, 2013.
- Zhou, Z.; Liu, X. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. **IEEE Trans. Knowledge Data Eng.**, v. 18, p. 63–77, 2006.
- Zhang, Y.; Bigham, J.; Ragland, D.; Chen, X. Investigating the associations between road network structure and non-motorist accidents. **Journal of Transport Geography**, v. 42, p. 34–47, 2015.
- Zhang, P.; Wang, J.; Li, X.; Li, M.; Di, Z.; Fana, Y. Clustering coefficient and community structure of bipartite networks, **Physica A: Statistical Mechanics and its Applications**, v. 387, n. 27, p. 6869–6875, 2008.
- Zhu, X.; Yuan, Y.; Hu, X.; Chiu, Y.; Ma, Y. A Bayesian Network model for contextual versus non-contextual driving behavior assessment. **Transportation Research Part C**, v. 81, pp. 172–187, 2017.
- Zhou, T.; Lü, L.; Zhang, Y. C. Predicting missing links via local information. **European Physical Journal B**, v. 71, n. 4, p. 623–630, 2009.

Apêndice A – Classificação taxonômica dos artigos selecionados na revisão sistemática da literatura.

Tabela 9 – Classificação dos trabalhos analisados

Autores	Contexto	Área de Concentração	Foco	Problema	Método de Pesquisa	Benefícios de desempenho	Número de objetivos	Posição da variável no modelo	Número de variáveis
Iranitalab e Khattak (2017)	A	S	FLO, MGL, K, ANN, SVM e NCC	PRE, AGR e CLA	QT, QA e ES	A e C	C	B	G
Nitsche <i>et al.</i> (2017)	A	S	AGR, CLA e VISIO	K e ASS	QT, QA e ES	A e C	C	B	B
Kumar, Toshniwal e Parida (2017)	B	S e P	K, LCC e ASS	AGR	QT, QA e ES	A e C	C	B	D
Nowakowska (2017)	A	S	RL, MB e BAL	CLA	QT, QA e ES	A e C	C	B	E
Moghaddam <i>et al.</i> (2017)	A	S	RBN, TET e EB	PRE e CLA	QT, QA e ES	A e C	C	B	C
Halim <i>et al.</i> (2016)	B	S	ANN, CART, K, PCA, FLO, GEN, SVM, SIT e SIMIO	PRE, CLA e AGR	RE	A, C e D	C	A, B e C	A
Saha, Alluri, Gan (2011)	A	I e S	FLO, HSM e EB	PRE, AGR	QT, QA e ES	C	C	B	C
Chen <i>et al.</i> (2016)	A	S	MGL, RBN e CART	PRE e CLA	QT, QA e ES	C	C	B	D
Huang <i>et al.</i> (2016)	A	S	ANN, SIG e TET	PRE	QT, QA e ES	C	C	B	E
Mujalli, López e Garach (2016)	A	S	CART, RBN, BN e BAL	CLA	QT, QA e ES	C	C	B	D
Yao, Loo e Yang (2016)	A	S	EB, MMC, FDK, SIG e M	AGR e CLA	QT, QA e ES	C	C	D	A
Costa <i>et al.</i> (2016)	A	S	MLG e RBN	PRE	QT, QA e ES	C	C	B	C
Kumar e Toshniwal (2015)	B	S	K, CART, RBN, TET e ASS	AGR	QT, QA e ES	C	C	B	D

Autores	Contexto	Área de Concentração	Foco		Método de Pesquisa	Benefícios de desempenho	Número de objetivos	Posição da variável no modelo	Número de variáveis	de
			Problema							
Chen, Guo e Wang (2015)	A	T e S	CLA, PRE e VISIO	SIT e FDK	QT, QA e ES	A e C	C	A e B	C	
Kwon, Rhee e Yoon, (2015)	A	S	CLA	RBN, RL e CART	QT, QA e ES	C	C	B	C	
Theofilatos e Yannis (2014)	A	S	PRE e CLA	SIT	RE	A e C	C	D	A	
De Oña <i>et al.</i> (2013)	A	S	CLA	MLG, BN e CART	QT, QA e ES	C	C	B	E	
Plug, X e Caulfield (2011)	A	S	CLA	SIG, M e FDK	QT, QA e ES	C	C	D	A	
Ahmed <i>et al.</i> (2011)	A	S	CLA	MB e PO	QT, QA e ES	A e C	C	A e B	E	
De Oña, Mujalli e Calvo (2011)	A	S	PRE e CLA	RL, PROB, TET e BN	QT, QA e ES	A e C	C	A e B	E	
Karlaftis e Vlahoganni (2011)	A	T, I, P, F, E, Q e S	PRE, CLA, AGR, DET	TET e ANN	QT, QA, CO, RE e LE	A, B e C	C	D	A	
Huang e Abdel-Aty (2010)	A	S	PRE	M, MB e MUL	QT, QA e ES	A e C	C	B	C	
Anderson (2009)	A	S	CLA	FDK, K e M	QT, QA e ES	A e C	C	B	C	
Depaire, Wets e Vanhoof (2008)	A	S	AGR	K e LCC	QT, QA e ES	A e C	C	A e B	F	
Li <i>et al.</i> (2008)	A	S	PRE	EB, MLG, SVM, BN e ANN	QT, QA e ES	A e C	C	A e B	B	
Xie, Lord e Zhang (2007)	A	S	PRE	RBN, ANN, MB e BN	QT, QA e ES	A e C	C	A e B	B	
Caliendo, Guida e Parisi (2007)	A	I e S	PRE e CLA	PCA, TET, CA, MLG e PO	QT, QA e ES	A e C	B	B	C	
Chang e Wang (2006)	A	S	CLA	CART e TET	QT, QA e ES	A e C	A	B	C	
Pande e Abdel-Aty (2006)	A	T e S	CLA	ANN, SIT e SIT	QT, QA e ES	A e C	C	A e B	C	
Delen, Sharda e Bessonov (2006)	A	I e S	CLA	ANN e DEP	QT, QA e ES	A e C	C	B	C	
Chang e Chen (2005)	A	S	CLA	RM e RBN	QT, QA e ES	A e C	C	B	C	
Chang (2005)	A	S	CLA	ANN e RBN	QT, QA e ES	A e C	C	B	C	

Autores	Contexto	Área de Concentração	Problema	Foco	Método de Pesquisa	Benefícios de desempenho	Número de objetivos	Posição da variável no modelo	Número variáveis	de
Sohn e Lee (2003)	A	S	CLA, AGR	ANN, K e PCA	QT, QA e ES	A e C	C	D	A	
Mussone, Ferrari e Oneta (1999)	A	S	PRE e CLA	ANN	QT, QA e ES	A e C	C	B	C	
Marhaviilas, KoulouriotiS e Gemeni, (2011)	A	S	CLA	TET	RE	A e C	C	D	A	
Teodorovi (2008)	A	T, I, P e Q	PRE e CLA	GEN	QT, QA e CO	A e C	C	D	A	
Thill (2000)	A	I, S	CLA	SIG	CO	C	C	D	G	
Mannering e Bhat (2014)	A	S	PRE	TET	RE	A e C	C	D	A	
Lin, Wang e Sadek (2014)	A	S	CLA, AGR e DET	NET, K, TET, ASS e FDK	QT, QA e ES	C	C	B	E	
Ting, Bin e Qi (2016)	A	T e S	PRE, AGR e DET	SIT e NET	QT, QA e CO	A, B e C	C	A e B	B	
Saha, Alluri e Gan (2015)	A		PRE, CLA e AGR	K, HSM, EB e FLO	QT, QA e ES	A e C	C	A e B	A	
Park e Abdel-Aty (2015)	A	S	CLA e PRE	ANN, RBN e HSM	QT, QA e ES	C	C	A e B	C	
Venkataraman, Ulfarsson e Shankar (2014)	A	S	PRE	TET e HSM	QT, QA e CO	A e C	C	D	A	
Young e Park (2013)	A	S	PRE	RBN, HSM e TEST	QT, QA e ES	A e C	C	A e B	C	

Apêndice B – Variáveis utilizadas nos trabalhos analisados.

Tabela 10 – Variáveis utilizadas em cada trabalho analisado.

Autores	Variáveis
Iranitalab e Khattak (2017)	tipo de rodovia, características da rodovia, tipo de pavimento, condição da pista, número de faixas, tipo de barreira, tipos de irregularidade do pavimento, perfil da rodovia, dispositivos de controle de tráfego, zonas de trabalho, condição da pista, número de interseções, , características dos motoristas, tipo de veículo, características dos acidentes, características de uso e ocupação do solo, tipo de propriedade do veículo (particular ou público), grupo da população, condições ambientais, condições de iluminação, condição de tempo, animais na rodovia, obstrução de visibilidade, detritos na pista, obstruções na rodovia
Nitsche <i>et al.</i> (2017)	o tipo de junção (tipo “T”, cruzamentos, múltiplo entroncamento, outros cruzamentos, rota privada ou dispositivos de acesso) e a gravidade dos acidentes (fatal, grave ou leve), com ou sem sinalização
Kumar, Toshniwal e Parida (2017)	nível de severidade, idade das vítimas, sexo, período do dia, mês, iluminação, área ao redor e tipo de acidente
Nowakowska (2017)	não específica
Moghaddam <i>et al.</i> (2017)	idade, sexo, nível educacional, preço do carro, área de residência, nível de renda, experiência de dirigir, exposição ao tráfego e tipos de personalidade.
Halim <i>et al.</i> (2016)	rodovia, veículo, via e meio ambiente
Saha, Alluri, Gan (2011)	tipo de área, classificação da área urbana/rural, número de faixas, canteiro central, largura da faixa, tipo de pavimento, Volume Médio Diário anual Total (AADT), comprimento do segmento, limite de velocidade, largura do acostamento.
Chen <i>et al.</i> (2016)	tipo de colisão, local do evento, condição de iluminação, condição climática, perfil da estrada, número de veículos envolvidos, função de estrada, envolvimento de motocicletas, envolvimento de veículos pesados, distância do local do acidente até a interseção mais próxima, condição do pavimento, número de faixas de rolagem, uso do cinto de segurança, envolvimento do motorista com álcool e dano ao veículo
Huang <i>et al.</i> (2016)	Acidentes por km, AADT, comprimento do segmento, número de faixas, largura média de cada faixa, velocidade limite, número de rampas, gradiente médio, raio médio de curvatura, largura do acostamento, barreiras, precipitação anual, número de paradas de ônibus, números de estacionamentos.
Mujalli, López e Garach (2016)	número de veículos envolvidos, tipo de acidente, padrão de colisão, sentido da via, número de faixas, traçado da rodovia, perfil, tipo de pavimento, condição da pista, condição do tempo, condição de iluminação, controle de tráfego (polícia, sinalizações na via, sinalizações luminosas, radar, câmeras), limite de velocidade.

Autores	Variáveis
Yao, Loo e Yang (2016)	não se aplica.
Costa <i>et al.</i> (2016)	volume de tráfego, distância percorrida, geometria, tipo de controle, etc.
Kumar e Toshniwal (2015)	número de vítimas, idade, sexo, hora do dia, mês, iluminação na estrada, tipo de rodovia, tipo de segmento, severidade do acidente, tipo de acidente e área ao redor
Chen, Guo e Wang (2015)	velocidade de tráfego, volume de tráfego, número de faixas, status da faixa, comunicações e sinalização viária.
Kwon, Rhee e Yoon, (2015)	comportamento do condutor, características do condutor, características do veículo, condição da pista, hora do dia, localização, características ambientais e tipo de acidente
Theofilatos e Yannis (2014)	não se aplica
De Oña <i>et al.</i> (2013)	tipo de acidente, idade, fatores atmosféricos, causa, dia, gênero, largura da pista, iluminação, mês, número de vítimas, número de ocupantes nos veículos, condição do pavimento, largura da faixa, sinalização horizontal, acostamento, distância de visibilidade, hora do dia e número de veículos envolvidos
Plug, X e Caulfield (2011)	não se aplica
Ahmed <i>et al.</i> (2011)	total de acidentes por segmento, comprimento do segmento, volume de tráfego, comprimento do deslocamento da viagem, estação do ano, inclinação do segmento, raio da curva, deflexão da curva, comprimento da curva, raio do comprimento com curva, número de faixas, largura média da faixa, acostamento, largura do acostamento, limite de velocidade, percentual de caminhões
De Oña, Mujalli e Calvo (2011)	tipo de acidente, idade, fatores atmosféricos, causa, dia, gênero, largura da faixa, iluminação, mês, número de vítimas, número de ocupantes nos veículos, acostamento, largura do pavimento, sinalização horizontal, sinalização vertical, distância de visibilidade, hora do dia e número de veículos envolvidos.
Karlaftis e Vlahoganni (2011)	integração de várias bases de dados.
Huang e Abdel-Aty (2010)	diferentes bases de dados de acidentes de tráfego
Anderson (2009)	Causas prováveis, número de pedestres, número de estações de metrô, número de paradas de ônibus, número de veículos, hora do dia, dia da semana, condição do tempo, número de acidentes com ciclistas, severidade.
Depaire, Wets e Vanhoof (2008)	tipo de veículo, idade, tipo de cruzamento, área construída, tipo de rodovia, canteiro central, limite de velocidade, faixas de velocidade, zonas críticas, infraestrutura, dia da semana, hora do dia, estação do ano, número de colisões com pedestres, níveis de segurança, chuva, condição da pista, tipo de acidente, posição do passageiro, número de passageiros, número de acidentes, eventos dinâmicos e consequência.
Li <i>et al.</i> (2008)	comprimento do segmento, largura do acostamento, volume de tráfego, número de faixas, largura da faixa

Autores	Variáveis
Xie, Lord e Zhang (2007)	comprimento do segmento, volume de tráfego, número de faixas, número de acidentes por ano, largura da faixa
Caliendo, Guida e Parisi (2007)	comprimento, curvatura, AADT, distância de visibilidade, coeficiente de atrito lateral, declive longitudinal e presença de entroncamento.
Chang e Wang (2006)	sexo, idade, localização do acidente, condição de sobriedade, tipo de veículo, causas prováveis e tipo de colisão
Pande e Abdel-Aty (2006)	velocidade média, velocidade máxima, local do acidente, tempo de congestionamento, número de faixas congestionadas, AADT.
Delen, Sharda e Bessonov (2006)	severidade, sexo, tipo de veículo, variáveis ambientais, informação do acidente e outras informações
Chang e Chen (2005)	frequência de acidentes por ano, grau da curva horizontal, grade vertical, volume de caminhões, volume de veículos reboque, volume de ônibus, fator de hora pico, número de dias com precipitação e precipitação anual
Chang (2005)	frequência de acidentes, comprimento da seção, raio da curva, grade vertical, volume de tráfego, percentual de caminhões, percentual de ônibus, número de dias de precipitação e volume de precipitação anual
Sohn e Lee (2003)	não se aplica
Mussone, Ferrari e Oneta (1999)	número de acidentes, local do acidente, características do condutor, gravidade da lesão, número de vítimas
Marhavidas, Koulouriotis e Gemeni, (2011)	não se aplica
Teodorovi (2008)	não se aplica
Thill (2000)	não se aplica
Mannering e Bhat (2014)	não se aplica
Lin, Wang e Sadek (2014)	estação do ano, dia da semana, hora do dia, velocidade do vento, condição climática, sentido da via, número de faixas, número de rampas, tipo de rampa, tipo de veículo, local do acidente, descrição do local do acidente, número de veículos envolvidos, tempo sem acidentes, índice de faixa bloqueada, número da faixa bloqueada, número de vítimas, capotamentos, congestionamento
Ting, Bin e Qi (2016)	dados simulados que incluem locais, velocidades e orientações dos veículos
Saha, Alluri e Gan (2015)	tipo de área, função da área, número de faixas, canteiro central, acostamento, fluxo médio anual de veículos, parâmetros do modelo de calibração, estacionamento, áreas comerciais, comprimento da calçada, etc.
Park e Abdel-Aty (2015)	volume de tráfego, características geométricas da estrada e outros fatores relacionados ao ambiente rodoviário

Autores	Variáveis
Venkataraman, Ulfarsson e Shankar (2014)	não se aplica
Young e Park (2013)	parâmetros de calibração do HSM

Apêndice C –Trabalhos selecionados.

Tabela 11 - Descrição dos trabalhos selecionados.

Autores	Descrição do Trabalho
Iranitalab e Khattak (2017)	<p>Compara o desempenho de quatro métodos estatísticos e de aprendizado de máquina, incluindo Multinomial Logit, vizinho mais próximo, máquina de vetores de suporte e florestas aleatórias, na previsão de acidentes de trânsito. Apresenta também uma abordagem baseada em custos de acidentes do tipo colisão pela comparação de métodos de previsão de severidade. Além disso, investiga os métodos de agrupamento de dados que incluem o K-means e classes latentes, sobre o desempenho de modelos de previsão de severidade de acidentes. A abordagem mostrou que a classificação usando o algoritmo do vizinho mais próximo teve o melhor desempenho de previsão em geral e em colisões mais graves. Os algoritmos de florestas aleatórias e suporte de vetores de máquina tiveram desempenhos próximos, enquanto que o modelo de regressão logística multinomial apresentou um desempenho mais fraco. O agrupamento de dados com o k-means não afetou os resultados de predição usando máquinas de vetores de suporte, mas melhorou o desempenho de previsão do modelo logística multinomial, do vizinho mais próximo e florestas aleatórias. Enquanto que o agrupamento por classes latentes apresentou uma melhoria de desempenho nas abordagens de regressão logística multinomial e florestas aleatórias, mas enfraqueceu o desempenho usando o classificador dos k-vizinhos mais próximos. A taxa geral de previsão utilizando a abordagem proposta, mostrou que negligenciar os custos do acidente pode levar ao erro de julgamento na escolha do método de previsão correto.</p>
Nitsche <i>et al.</i> (2017)	<p>Apresenta um método de análise de dados, incluindo a preparação, a análise e visualização de acidentes de carro, para identificar os cenários críticos pré-colisão em diferentes padrões cruzamentos, para testar a segurança de sistemas de acionamento automatizados. Os resultados apoiam as descobertas existentes sobre acidentes em cruzamentos rodoviários e fornecem situações de referência para testes de desempenho de segurança, a fim de reduzir as possíveis combinações de parâmetros numéricos.</p>
Kumar, Toshniwal e Parida (2017)	<p>Aplica e avalia o desempenho de técnicas de agrupamento LCC e k-medóides em dados de acidentes rodoviários. Foram geradas regras de associação a partir dos agrupamentos obtidos pelas diferentes abordagens. Os resultados revelam que as técnicas são compatíveis e apresentam desempenhos semelhantes. No entanto, ambas são eficientes na redução da heterogeneidade dos dados. As regras de decisão obtidas revelam algumas informações que podem ser utilizadas no desenvolvimento de políticas públicas e para prevenir e superar as taxas de acidentes de trânsito.</p>
Nowakowska (2017)	<p>Discute o desenvolvimento do modelo Bayesiano de regressão logística na classificação da severidade dos acidentes rodoviária, com base na definição de informações de verossimilhança na forma de conjuntos de dados de treinamento desequilibrados e balanceados. Os modelos obtidos a partir do conjunto de dados de treinamento equilibrado têm uma melhor qualidade de classificação do que os obtidos a partir do conjunto de dados de treinamento desequilibrado.</p>

Autores	Descrição do Trabalho
Moghaddam <i>et al.</i> (2017)	<p>Desenvolveu modelos de previsão de acidentes, considerando a estrutura de erros do modelo binomial negativo e as características de vários condutores. O estudo permitiu correlacionar o nível de escolaridade dos condutores, com sexo, idade, com o tipo e valor do veículo. Os resultados indicaram que a idade, sexo, nível de escolaridade, anos de direção ativa e, especialmente, a exposição ao tráfego tiveram um efeito considerável sobre os acidentes com vítimas, enquanto não houve efeito discernível no nível de renda, tipo de personalidade e área de residência.</p> <p>Apresenta um levantamento abrangente de várias técnicas de I. A. usadas na literatura para prever um acidente e também estudar padrões de condução de risco. O estudo revela que a ANN, SVM e os algoritmos genéticos têm sido as técnicas de I. A. mais utilizadas na predição de acidentes/ padrão de condução de risco. Os melhores resultados obtidos foram por meio das modelagens usando ANN e árvores de decisão. No entanto, essa precisão depende do tipo de dados, da quantidade de dados e das características levadas em consideração pela técnica de predição de acidentes. Destaca que a maioria dos estudos utilizam dados de simulação ou simuladores para testar as diferentes abordagens, bem como o auxílio de classificadores principalmente para modelar características relacionadas ao comportamento do condutor. No entanto a precisão de predição é o aspecto mais crítico em segurança viária e as técnicas de I. A. tem contribuído para melhorar isso.</p>
Halim <i>et al.</i> (2016)	<p>Apresenta um método de priorização das variáveis de calibração do HSM com base em seu impacto nas previsões de acidentes do tipo colisão usando algoritmos de florestas aleatórias e clusterização de dados. A abordagem eficiente e permitiu identificar que as variáveis mais influentes aos acidentes de tráfego do tipo colisão foram o volume de tráfego, bem como a densidade de áreas comerciais e a presença de áreas residenciais.</p>
Saha, Alluri, Gan (2011)	<p>Aplica um classificador híbrido utilizando os algoritmos Naive Bayes e árvore de decisão em um conjunto de dados de acidentes de trânsito do tipo colisão traseira. A abordagem não paramétrica permitiu selecionar os fatores que contribuíram a ocorrência dos acidentes e a gravidade da lesão dos motoristas. O classificador híbrido forneceu um resultado de classificação razoável, indicado por várias medidas de desempenho, como medida F, Curva ROC e AUC, bem como pelos resultados de comparação de desempenho do modelo com vários modelos populares.</p>
Chen <i>et al.</i> (2016)	<p>Desenvolve uma função de base radial usando a abordagem de ANN otimizada, visando aproximar as relações não lineares entre frequência de acidente e os fatores de riscos. O desempenho de previsão do modelo é comparado em relação aos obtidos com regressão binomial negativa (NB) e redes neurais artificiais clássicas (ANN), para a previsão de frequências de acidentes do tipo colisão em segmentos rodoviários. Os resultados indicam que a abordagem utilizando ANN com funções radiais tem melhor ajuste e desempenho de previsão do que os modelos NB e ANN.</p>
Huang <i>et al.</i> (2016)	<p>Investiga a possibilidade de se utilizar técnicas de amostragem em dados de acidentes de trânsito desequilibrados na etapa prévia do processo de classificação utilizando diferentes classificadores de Bayes, a fim de desenvolver modelos de previsão do nível de severidade de um acidente de trânsito. Utiliza dados obtidos para estradas urbanas e suburbanas. Os autores concluem que o balanceamento do conjunto de dados melhora o desempenho dos classificadores de Bayes no processo de classificação das lesões graves e fatais, bem como da classe majoritária de dados (lesões leves, etc.). Por outro lado, os modelos de Bayes desenvolvidos a partir do conjunto de dados desequilibrados foi incapaz de determinar quais valores de variáveis foram os que mais influenciaram o resultado de uma lesão grave ou fatal em um acidente.</p>
Mujalli, López e Garach (2016)	

Autores	Descrição do Trabalho
Yao, Loo e Yang (2016)	<p>Fornece uma análise sistemática dos principais avanços no GIS aplicado para estudar colisões de tráfego no espaço desde meados da década de 1970. Os autores destacam que o padrão espacial das colisões deve ser explorado de maneira científica. Destacam que as ferramentas de GIS fornecem uma visualização simples de seções de estradas perigosas e apresentar a capacidade de analisar a dependência espacial e as interações espaciais com outros fatores geográficos. Desenvolve um modelo binomial negativo e generalizados para a previsão de acidentes do tipo colisão em junções de três pernas, com a finalidade de identificar os fatores que contribuíram a ocorrência dos acidentes. Para tanto foram consideradas as características geométricas de cada entroncamento e de sua área de influência. Os resultados confirmam que o uso de métodos de calibração que consideram a heterogeneidade não observada leva à obtenção de modelos que são mais adequados para a estimar a frequência de acidentes observados.</p>
Costa <i>et al.</i> (2016)	<p>Usou técnicas de agrupamento e classificação de dados, visando otimizar o processo de classificação a partir da segmentação e redução prévia da heterogeneidade dos dados de acidentes. As regras de associação geradas para cada cluster identificaram as diferentes circunstâncias propensas a acidentes em cada cluster. Os resultados mostraram que realizar a análise de cluster como uma tarefa preliminar da classificação e obtenção de regras de associação, pode identificar relacionamentos importantes que não são verificados quando o conjunto de dados inteiro é analisado.</p>
Kumar e Toshniwal (2015)	<p>Apresenta conceitos básicos relacionados à visualização de dados de tráfego. Fornece uma visão geral das técnicas de processamento de dados relacionadas e resume os métodos existentes para descrever as propriedades temporais, espaciais, numéricas e categóricas dos dados de tráfego, com base na utilização de sistemas inteligentes e na função de densidade de kernel.</p>
Chen, Guo e Wang (2015)	<p>Apresenta os classificadores Naive Bayes e de árvore de decisão na classificação dos fatores de risco em ordem de importância que afetam o nível de severidade dos acidentes. Os autores destacam que as metodologias usadas são prontamente aplicáveis aos outros conjuntos de dados existentes para avaliar a dependência entre os fatores de risco e para entender sua importância relativa na severidade dos acidentes.</p>
Kwon, Rhee e Yoon, (2015)	<p>Fornece uma revisão do efeito das características do tráfego e do clima na segurança rodoviária. Identifica as lacunas e discute as necessidades de pesquisas adicionais. Ressaltam que o fluxo de tráfego parece ter uma relação não linear com as taxas de acidentes. Os resultados sugerem que baixa visibilidade, alta precipitação e variação de velocidade aumentam a probabilidade de colisões.</p>
Theofilatos e Yannis (2014)	<p>Mostra a possibilidade de utilização de Redes Bayesianas (BNs) para classificar acidentes de trânsito de acordo com a gravidade da lesão. As BNs são capazes de fazer previsões e classificações sem a necessidade da utilização de pré-suposições entre o conjunto de variáveis utilizado. Além disso, são usadas para fazer representações gráficas de sistemas complexos com componentes inter-relacionados. Este estudo destacou a possibilidade de utilizar as BNs para modelar acidentes de trânsito, embora o foco não tenha sido a análise de desempenho do classificador bayesiano.</p>
De Oña <i>et al.</i> (2013)	

Autores	Descrição do Trabalho
Plug, X e Caulfield (2011)	<p>Aplicou técnicas espaciais, temporais e espaço-temporais baseadas em SIG e na função de estimativa de densidade de Kernel para investigar padrões de acidentes do tipo colisão. Os resultados obtidos revelam não apenas padrões de colisão de veículos em uma determinada área e período de tempo, mas também demonstram a existência de processos subjacentes aos padrões de colisão. Os hotspot críticos determinados temporalmente pela função de densidade de kernel (KDE -Kernel density estimation) pode mostrar as mudanças de padrões durante um certo período de tempo. Ao cruzar esses padrões e fatores de risco previamente conhecidos, como velocidade e fadiga, alguns fatores significativos podem ser identificados para entender as razões por trás de certos padrões de acidentes.</p>
Ahmed <i>et al.</i> (2011)	<p>Este artigo examina os efeitos da geometria da estrada na ocorrência de acidentes em uma autoestrada com terreno montanhoso e em condições climáticas adversas. A estimativa dos coeficientes do modelo hierárquico bayesiano indica que a geometria da pista está significativamente associada ao risco de colisão, enquanto que o aumento do número de faixas pode estar associado a diminuição deste tipo de acidente. Os resultados identificaram tendências claras associadas ao efeito da inclinação, quanto mais íngreme a inclinação de um segmento, maior será o risco de colisão. Este risco se acentuada em temporadas de neve, representando níveis de risco significativamente maiores do que na estação seca. Além disso, os modelos bayesianos apresentaram um melhor desempenho quando comparados aos de Poisson no que se refere a modelagem de efeitos aleatórios.</p>
De Oña, Mujalli e Calvo (2011)	<p>Apresenta uma análise de acidentes em rodovias rurais visando detectar os fatores chaves na ocorrência de trânsito usando BNs. Os resultados indicam que as BNs são eficientes na classificação da severidade dos acidentes e na identificação dos fatores correlacionados à ocorrência.</p>
Karlaftis e Vlahoganni (2011)	<p>Fornece uma avaliação do desempenho da aplicação de abordagens de modelagem por métodos estatísticos e de redes neurais em pesquisas de transporte . Os resultados sugerem um melhor desempenho das redes neurais no tratamento de dados complexos e não-lineares (multidimensionais), com o tratamento de outliers, dados ausentes ou ruidosos, e da multicolinearidade. Os autores destacam o maior poder de explicação das técnicas estatísticas tradicionais, embora limitadas a problemas de pequena escala. A recomendação é as técnicas sejam usadas de forma sinérgica em diferentes estágios da modelagem de dados.</p>
Huang e Abdel-Aty (2010)	<p>O estudo demonstra que os modelos espaciais hierárquicos obtidos pela abordagem de dados multinível e pela análise bayesiana, fornecem uma melhor representação dos processos estocásticos sob os dados de segurança observados com estruturas multiníveis.</p>
Anderson (2009)	<p>Este trabalho propõe uma abordagem metodológica que se baseia em GIS e estimativa de densidade de kernel para criar uma unidade espacial básica de um hotspot. Os resultados estatisticamente e agrupados utilizando a técnica de clustering de acordo com os dados do atributo do acidente e da área circundante. Este artigo agrega valor significativo à investigação sobre a delimitação de pontos críticos de acidentes rodoviários e revela a natureza complexa de como medimos os pontos críticos de acidentes rodoviários.</p>

Autores	Descrição do Trabalho
Depaire, Wets e Vanhoof (2008)	<p>Examina a eficácia da técnica de agrupamento por classes latentes na identificação de acidentes de tráfego homogêneos. Os autores destacam que a segmentação dos acidentes em cluster é eficiente para análise de grandes e heterogêneas bases de dados, contribuindo a análise da gravidade das lesões em diferentes tipos de acidentes.</p>
Li <i>et al.</i> (2008)	<p>O objetivo deste estudo foi avaliar a aplicação de modelos SVM na previsão de colisões de veículos. Para tanto foram desenvolvidos e comparados modelos de NB regressão e SVM com base em dados coletados em estrada rurais. Amostras de diferentes tamanhos foram usadas para a comparação dos modelos. Os modelos também foram comparados aos modelos de ANN. O estudo mostra que os modelos de SVM predizem dados de colisão com mais eficácia e precisão do que modelos tradicionais de NB, além de serem mais rápidos de implementar quando comparados as ANN.</p>
Xie, Lord e Zhang (2007)	<p>Avalia a aplicação de modelos obtidos com ANN e redes Bayesianas na previsão de colisões de veículos em redes de transporte. Os modelos apresentam um bom desempenho quando comparados aos modelos de regressão. Contudo, os autores descartam a limitação do poder de ajuste da abordagem considerando os dados analisados.</p> <p>Apresenta modelos de previsão de acidentes focados em estradas rurais multilane com base em acidentes observados e utilizando Modelos Lineares Generalizados (MLG). Os resultados foram estimados utilizando o método da Máxima Verossimilhança. Os modelos de previsão foram separados considerando o total de acidentes com colisão e o número de acidentes fatais por colisão. O efeito da precipitação pluviométrica foi analisado com base em dados e suposições de precipitação por hora sobre o tempo de secagem. Os resultados indicam que o pavimento molhado aumenta significativamente o número de colisões. Os modelos de acidentes propostos indicam que a maior parte dos acidentes estão relacionados ao fluxo de tráfego, a geometria da estrada e a precipitação. Os modelos podem ser refinados com a utilização de variáveis relacionadas ao comportamento humano contribuindo significativamente ao aumento da variância total explicada.</p>
Caliendo, Guida e Parisi (2007)	<p>Um modelo baseado em árvore (CART) foi desenvolvido para estabelecer a relação entre gravidade da lesão, as características do veículo, as variáveis rodoviárias, ambientais e do acidente. Os resultados indicam que a variável mais importante associada à gravidade do acidente é o tipo de veículo. Pedestres, motociclistas e ciclistas são identificados como tendo os maiores riscos de serem feridos do que outros tipos de condutores de veículos em acidentes de trânsito. Este estudo demonstra que a classificação e árvores de regressão é uma boa alternativa para analisar a gravidade da lesão em acidente de trânsito.</p>
Chang e Wang (2006)	<p>Apresenta modelos de classificação baseados na abordagem de mineração de dados. Os autores identificam que os algoritmos de classificação atingem precisão satisfatória sobre o conjunto de dados de validação. Os resultados indicam que esses modelos podem ser aplicados para identificar condições de tráfego em tempo real e identificar acidentes relacionados à mudança de faixa.</p>
Pande e Abdel-Aty (2006)	<p>Este estudo utiliza uma série de redes neurais artificiais para modelar relações potencialmente não-lineares entre os níveis de gravidade das lesões e os fatores relacionados a acidentes. O trabalho apresenta a análise de sensibilidade dos modelos de rede neural obtidos para identificar a importância priorizada dos fatores relacionados a acidentes e a influência destes fatores sobre os diferentes níveis de gravidade de lesões</p>
Delen, Sharda e Bessonov (2006)	

Autores	Descrição do Trabalho
Chang e Chen (2005)	Compara o desempenho da previsão entre os modelos de regressão binomial negativa e CART . O estudo demonstra que CART apresenta um desempenho melhor e pode ser utilizada como método alternativo a análise de frequências de acidentes em autoestradas.
Chang (2005)	Compara o desempenho de previsão entre o modelo de regressão binomial negativo e o modelo de ANN. Demonstra que a ANN é um método alternativo consistente para a análise de frequência de acidentes em rodovias.
Sohn e Lee (2003)	Foi feita uma fusão entre técnicas de inteligência artificial como PCA, Clustering e ANN visando melhorar a precisão da classificação da gravidade dos acidentes de trânsito. Destacam a boa performance da classificação a partir de agrupamentos obtidos por clustering.
Mussone, Ferrari e Oneta (1999)	Descreve um modelo alternativo baseado em ANN visando a análise de acidentes veiculares urbanos. Os autores identificaram que quando maior a complexidade do padrão da interseção maior será o índice de acidentes. O maior índice de acidentes do tipo atropelamento foi detectado em cruzamentos não sinalizados durante a noite. Os autores destacam a eficácia da ANN na classificação dos fatores críticos que causam os acidentes
Marhaviilas, KoulouriotiS e Gemeni, (2011)	Classifica e categoriza os principais métodos e técnicas de análise de risco e avaliação de risco por meio da avaliação de artigos científicos em diversas áreas do conhecimento. A revisão da literatura científica mostrou que as técnicas de análise e avaliação de risco são classificadas em três categorias principais: as técnicas qualitativas, as quantitativas e as técnicas híbridas (qualitativas, quantitativas, semi-quantitativas). A análise estatística mostra que os métodos quantitativos apresentam a maior frequência relativa em aplicações enquanto a qualitativa é menor. Além disso, os métodos híbridos permanecem constantemente em um nível muito baixo durante todo o período de processamento.
Teodorovi (2008)	Aborda a aplicação de algoritmos genéticos baseado em otimização de colônias de formigas em Engenharia de Transportes como roteamento e programação de veículos. Os Algoritmos genéticos representam técnicas de busca baseadas na mecânica da seleção da natureza, sendo usadas na solução de problemas complexos de otimização combinatória. As rotas pesquisas pelas formigas representam as soluções de otimização do problema pesquisado. Os autores descartam que os algoritmos genéticos representam uma técnica promissora capaz de resolver problemas complexos de engenharia de tráfego e planejamento de transportes.
Thill (2000)	Apresenta uma pesquisa direcionada a melhora das abordagens de SIG, com a finalidade de permitir toda a gama de capacidades necessárias na pesquisa e gestão de transportes.
Mannering e Bhat (2014)	Fornece uma orientação na definição desses desafios e oportunidades, revisando primeiro a evolução das aplicações metodológicas e os dados disponíveis na pesquisa de acidentes rodoviários. É mostrado que novas metodologias podem abordar questões complexas relacionadas à heterogeneidade não observada, compensação de risco, correlações espaciais e temporais, e têm o potencial de expandir significativamente a compreensão dos muitos fatores que afetam a probabilidade e a gravidade (em termos de danos pessoais) de acidentes rodoviários. Isso, por sua vez, pode levar a contramedidas de segurança mais eficazes que podem reduzir substancialmente lesões e fatalidades relacionadas à rodovia.

Autores	Descrição do Trabalho
Lin, Wang e Sadek (2014)	<p>Explora o potencial em se utilizar um algoritmo de detecção de comunidade com otimização de modularidade e oito algoritmos de aprendizado de regras de associação, para identificar características importantes de acidentes. O algoritmo de detecção da comunidade foi bastante eficaz na identificação de clusters com características discerníveis; o agrupamento ajudou a desvendar relações e fatores causadores de acidentes que permaneceu escondido quando a análise foi realizada em todo o conjunto de dados. Especificamente, quando o algoritmo de regra de associação foi executado em todo o conjunto de dados neste estudo, o insight obtido foi muito limitado em comparação com o ganho da execução da análise nos clusters. Propõe um método de aviso de congestionamento com base na Internet de veículos (IOV) e na descoberta de redes complexas pela detecção de comunidades comunidade. As comunidades nesta rede complexa são descobertas pelo rápido algoritmo Newman (FN - Fast Newman), e os avisos de congestionamento são gerados de acordo com as comunidades selecionadas por escala e densidade. Este método pode detectar a tendência de agregação de tráfego e fornecer avisos antes que ocorra o congestionamento. As simulações mostram que o método proposto neste artigo é eficaz e praticável, e torna possível agir antes do congestionamento do tráfego.</p>
Ting, Bin e Qi (2016)	<p>Este estudo visa priorizar as variáveis do HSM para fins de calibração determinando sua influência nas previsões de colisão. Foi aplicado clustering para agrupar as variáveis e o algoritmo de florestas aleatórias para determinar o impacto das variáveis recomendadas pelo HSM em previsões de colisão em vias artérias urbanas e suburbanas na Flórida. Os modelos obtidos identificaram a variável fluxo de tráfego como sendo a mais influente. Os autores também identificaram que o aumento da densidade de áreas comerciais influencia diretamente na ocorrência de acidentes do tipo colisão.</p>
Saha, Alluri e Gan (2015)	<p>Investiga o potencial da aplicação de ANN no desenvolvimento de funções globais de segurança viária, para prever a frequência de acidentes em diferentes rodovias, utilizando o máximo de variáveis possíveis no estudo. Os autores destacam a capacidade de generalização das ANN e sua capacidade de reduzir o overfitting quando há poucos dados de entrada.</p>
Park e Abdel-Aty (2015)	<p>Apresenta uma aplicação da estrutura do HSM para avaliação da segurança no trânsito. Realizar uma comparação de desempenho de métodos tradicionais univariados, multivariados e hierárquicos bayesianos modelados com as variáveis do HSM. Os autores evidenciam que os modelos apresentam um melhor desempenho na predição quando são utilizadas as variáveis recomendadas no HSM.</p>
Venkataraman, Ulfarsson e Shankar (2014)	<p>Compara o desempenho das funções de desempenho de segurança (SPF - Safety Performance Functions) HSM com parâmetros calibrados e não calibrados. O HSM recomenda a utilização de parâmetros calibrados para as SPFs, de forma a refletir com precisão as características exclusivas de uma determinada região de estudo e visando um melhor ajuste dos resultados. O estudo utilizou testes estatísticos de adequação de ajuste e parcelas residuais cumulativas. As SPFs calibradas fornecem o melhor ajuste aos dados e foram melhores na predição de acidentes em interseções com 3 e 4 pernas.</p>
Young e Park (2013)	

Apêndice D

Tabela D: Valores dos VDM_{total} para o período de 2009 a 2016.

O Apêndice apresenta os valores dos dados obtidos durante a pesquisa, demonstrados nas tabelas a seguir. Ressalta-se que a soma do VDMA de passeio com o comercial não será igual ao total, pois estes valores são preditos separadamente.

Tabela D1.IV: VDMA previsto dos anos de 2009 e 2010.

Sentido	Km inicial	Km final	2009		2010			
			VDMA Passeio	VDMA Comercial(x2)	VDMA Total	VDMA Passeio	VDMA Comercial(x2)	VDMA Total
Norte	125,0	128,0	14.550	11.850	19.950	14.730	11.500	21.350
Sul	125,0	128,0	14.480	11.690	16.800	14.790	11.320	20.200
Norte	128,0	135,0	39.750	15.050	54.750	40.380	14.890	55.250
Sul	128,0	135,0	36.200	15.000	51.800	36.800	14.300	51.700
Norte	135,0	139,0	75.900	20.700	96.000	66.250	18.750	85.000
Sul	135,0	139,0	72.500	18.500	91.100	64.900	17.100	82.000
Norte	139,0	145,5	43.000	18.700	54.100	43.400	18.350	57.050
Sul	139,0	145,5	46.000	17.100	63.150	45.600	17.010	62.650

Tabela D2.IV: VDMA previsto dos anos de 2013 e 2014.

Sentido	Km inicial	Km final	2013		2014			
			VDMA Passeio	VDMA Comercial(x2)	VDMA Total	VDMA Passeio	VDMA Comercial(x2)	VDMA Total
Norte	125,0	128,0	15.435	10.796	26.231	16.167	10.914	27.081
Sul	125,0	128,0	15.803	10.703	26.506	16.694	10.567	27.260
Norte	128,0	135,0	42.645	14.696	57.341	42.944	14.490	57.434
Sul	128,0	135,0	38.642	11.749	50.391	40.022	13.986	54.008
Norte	135,0	139,0	44.239	13.944	58.184	19.375	9.174	28.548
Sul	135,0	139,0	47.440	14.154	61.594	27.570	11.449	39.019
Norte	139,0	145,5	44.021	15.701	59.722	45.639	17.950	63.589
Sul	139,0	145,5	43.154	15.638	58.792	44.551	18.520	63.070

Tabela D3.IV: VDMA previsto dos anos de 2015 e 2016.

Sentido	2015				2016			
	Km inicial	Km final	VDMA Passeio	VDMA Comercial(x2)	VDMA Total	VDMA Passeio	VDMA Comercial(x2)	VDMA Total
Norte	125,0	128,0	16.234	10.131	26.365	15.181	9.244	24.425
Sul	125,0	128,0	16.842	9.841	26.683	16.071	9.197	25.268
Norte	128,0	135,0	43.185	13.831	57.016	44.000	13.940	58.000
Sul	128,0	135,0	39.109	8.133	47.242	41.517	11.241	52.758
Norte	135,0	139,0	18.992	8.615	27.607	14.445	7.245	21.790
Sul	135,0	139,0	21.856	9.996	31.851	19.228	8.589	27.817
Norte	139,0	145,5	46.489	17.007	63.495	44.821	15.811	60.632
Sul	139,0	145,5	43.089	16.784	59.873	43.333	15.907	59.240

Apêndice E - Banco de dados com as variáveis adquiridas pelo protocolo de pesquisa

Tabela E.1 - Descrição das variáveis analisadas no banco de dados da rodovia SP-065.

Variáveis	Tipo	Descrição
Dia	Categórica	1. Segunda-feira; 2. Terça-feira; 3. Quarta-feira; 4. Quinta-feira; 5. Sexta-feira; 6. Sábado e 7. Domingo
Mês	Categórica	1. Janeiro; 2. Fevereiro; 3. Março; 4. Abril 5. Maio; 6. Junho; 7. Julho; 8. Agosto; 9. Setembro; 10. Outubro; 11. Novembro e 12. Dezembro
Ano	Numérica	2009 a 2016
Severidade	Categórica	0. Acidente sem vítima; 1. Acidente com vítima e 2. Acidente com vítima fatal
Tipo de acidente	Categórica	1. Colisão traseira; 2. Colisão frontal; 3. Colisão transversal; 4. Colisão lateral; 5. Engavetamento; 6. Capotamento; 7. Atropelamento; 8. Tombamento; 9. Choque e 10. Outros
Causa provável	Categórica	1. Motorista; 2. Veículo; 3. Via/meio ambiente e 4. Outros
Período	Categórica	0. Manhã; 1. Tarde e 2. Noite
KM	Numérica	Km 125 ao km 145,500
Sentido	Categórica	1. Norte e 2. Sul
Local do Acidente	Categórica	1. Acostamento; 2. Faixa de Rolamento; 3. Canteiro Central; 4. Alça de entrada; 5. Alça de saída; 6. Faixa da Direita e 7. Faixa da Esquerda
Vítimas Fatais	Categórica	0. Ocorrência sem VF e 1. Ocorrência com VF
Vítimas não fatais	Categórica	0. Ocorrência sem VNF e 1. Ocorrência com VF
Vítimas com danos materiais	Categórica	0. Ocorrência sem VM e 1. Ocorrência com VM
Condição de visibilidade	Categórica	1. Boa; 2. Parcial e 3. Ruim
Condição especial	Categórica	1. Poça d'água; 2. Óleo; 3. Fumaça e 4. Outros
Condição do tempo	Categórica	1. Bom; 2. Chuva; 3. Nublado; 4. Neblina e 5. Garoa
Condição da pista	Categórica	1. Seca; 2. Molhada e 3. Oleosa
Traçado da pista	Categórica	1. Reta e 2. Curva
Perfil da pista	Categórica	1. Em nível; 2. Aclive e 3. Declive
Inclinação da pista	Numérica	Calculada em porcentagem (%)
Acostamento	Categórica	0. Sim e 1. Não
Largura do Acostamento	Numérica	Fornecida em metros
Luminosidade	Categórica	0. Sim e 1. Não
Radar	Categórica	0. Sim e 1. Não
Trevo	Categórica	0. Sim e 1. Não
Ramo	Categórica	0. Sim e 1. Não
Velocidade máxima	Numérica	Fornecida em km/h
Sinalização de obras	Categórica	1. Não Existe; 2. Existe Bem Sinalizada e 3. Existe Mal Sinalizada
Sinalização Horizontal	Categórica	1. Existe e 2. Não existe
Sinalização vertical	Categórica	1. Existe e 2. Não existe
Latitude	Numérica	Fornecida em metros (UTM)

Variáveis	Tipo	Descrição
Longitude	Numérica	Fornecida em metros (UTM)
Número de faixas	Numérica	Quantificada por trecho
Largura da faixa crítica	Numérica	Fornecida em metros
Canteiro Central	Categórica	0. Sim e 1. Não
Largura do Canteiro Central	Numérica	Fornecida em metros
Viaduto	Categórica	0. Sim e 1. Não
Ponte	Categórica	0. Sim e 1. Não
Barreiras de contenção	Categórica	0. Sem barreiras; 1. Com barreiras e 2. Parcialmente ou misto
Volume médio diário de veículo de passeios	Numérica	14445 a 16234
VDM diário de veículos comerciais	Numérica	44821 a 75900
VDMA total	Numérica	14445 a 75900
Intervenção	Categórica	0. Concluída; 1. não houve e 2. Em andamento

Apêndice F - Análise da estrutura topológica com PCA dos agrupamentos obtidos pelos métodos de clustering e redes complexas homogêneas

Neste Apêndice são apresentados os gráficos elaborados a partir da técnica de PCA e da construção de grafos, com a finalidade de avaliar a estrutura topológica dos agrupamentos obtidos no espaço. Nos gráficos apresentados neste apêndice, as diferentes cores apresentam as observações atribuídas a diferentes agrupamentos.

Nos gráficos obtidos a partir da técnica de PCA cada dimensão representa as 3 componentes principais mais representativas. Os valores dos eixos de cada dimensão correspondem aos autovetores obtidos. Os sinais positivos e negativos indicam que as relações entre as observações atribuídas a um agrupamento são, respectivamente, diretamente ou inversamente proporcionais.

A Figura F.1 e a Figura F.2 apresentam, respectivamente, a projeção espacial dos registros de acidentes com VM no período de antes (2009-2012) e depois (2012-2016). Ao aplicarmos os algoritmos de particionamento k – means e Fastgreedy, nota-se que os mesmos convergem, em geral, com diferentes valores de agrupamento. A visualização dos agrupamentos projetados pela técnica PCA em 3 dimensões, sugerem a formação de grupos 4 agrupamentos no cenário de antes (2009-2012) e de 6 agrupamentos no cenário de depois (2012-2016).

A Figura F.3 e a Figura F.4 apresentam, respectivamente, a estrutura topológica 3D dos agrupamentos obtidos com o k – means e Fastgreedy usando PCA e os registros de acidentes com VM para o período de antes (2009-2012). Para o período de depois (2012-2016), foram obtidos para ambos os algoritmos 5 agrupamentos de ocorrências com VM, conforme ilustram a Figura F.5 e a Figura F.6.

Figura F.1 - Projeção 3D usando PCA com VM de 2009-2012.

Figura F.2 - Projeção PCA (3D) com VM de 2012-2016.

Figura F.3 - Agrupamentos de VM pelo k – means usando PCA de 2009-2012.

Figura F.4 - Agrupamentos de VM pelo Fastgreedy usando PCA de 2009-2012.

Figura F.5 - Agrupamentos com VM pelo k – means usando PCA de 2012-2016.

Figura F.6 - Agrupamentos com VM pelo Fastgreedy usando PCA de 2012-2016.

Para os registros de VNF no período de 2009-2012 verifica-se, com base na Figura F.7, que a estrutura do banco de dados sugere que estes dados possam ser particionados em 5 grupos. Para o período de depois (2012-2016), os registros com VNF sugerem que os dados possam ser particionados em 6 grupos, conforme ilustra a Figura F.8.

A Figura F.9 e a Figura F.10 apresentam, respectivamente, os agrupamentos obtidos com o k – means e Fastgreedy VNF para de 2009-2012. Para o período de depois (2012-2016) com VNF, o particionamento destes dados resultou em 4 agrupamentos tanto para o k – means (Figura F.11) quanto para o Fastgreedy (Figura F.12). Os resultados indicaram diferenças sutis entre os agrupamentos obtidos por ambos os algoritmos.

Figura F.7 - Projeção 3D usando PCA com VNF de 2009-2012.

Figura F.8 - Projeção 3D usando PCA com VNF de 2012-2016.

Figura F.9 - Agrupamentos com VNF pelo k – means usando PCA de 2009-2012.

Figura F.10 - Agrupamentos com VNF pelo Fastgreedy usando PCA de 2009-2012.

Figura F.11 - Agrupamentos com VNF pelo k – means usando PCA de 2009-2012.

Figura F.12 - Agrupamentos com VNF pelo Fastgreedy usando PCA de 2012-2016.

Os registros com VF para o cenário de antes (2009-2012) depois (2012-2016) ilustrados, respectivamente, pelas Figura F.13 e Figura F.14, sugerem que os dados não apresentam estrutura de agrupamentos. Tal evidência se reflete quando se analisa os

resultados dos agrupamentos obtidos com k – means (Figura F.15) e Fastgreedy (Figura F.16) para o período de antes. E para o período de depois pelas Figura F.17 e Figura F.18.

Figura F.13 - Projeção 3D usando PCA com VF de 2009-2012.

Figura F.14 - Projeção 3D usando PCA com VF de 2012-2016.

Figura F.15 - Agrupamentos com VF pelo k – means usando PCA de 2009-2012.

Figura F.16 - Agrupamentos com VF pelo Fastgreedy usando PCA de 2009-2012.

Figura F.17 - Agrupamentos com VF pelo k – means usando PCA de 2012-2016.

Figura F.18 - Agrupamentos com VF pelo Fastgreedy usando PCA de 2009-2012.

Apêndice G - Análise da estrutura topológica por meio de grafos dos agrupamentos obtidos pelos métodos de clustering e redes complexas homogêneas

Neste Apêndice são apresentados os grafos obtidos a partir das estruturas ótimas clustering e de redes detectadas, respectivamente, na seção 4.2 e 4.3. Nos grafos as diferentes cores apresentam as observações atribuídas aos diferentes agrupamentos resultantes dos algoritmos k – means e Fastgreedy.

A Figura G.1 e a Figura G.2 apresentam os grafos obtidos a partir dos bancos de dados de acidentes com VM, respectivamente para o cenário de antes (2009-2012) e depois (2012-2016). E a Figura G.3 e a Figura G.4 apresentam, respectivamente, os agrupamentos obtidos pelos métodos k – means e Fastgreedy.

Figura G.1 – Grafo obtido com VM de 2009-2012

Figura G.2 – Grafo obtido com VM de 2012-2016.

Figura G.3 – Agrupamentos k – means com VM de 2009-2012.

Figura G.4 – Agrupamentos Fastgreedy com VM de 2009-2012.

A análise da estrutura topológica do grafo ilustrado pela Figura G.3 sugere que o método k – means é mais propenso a formação de sub-agrupamentos, enquanto que o grafo obtido a partir dos agrupamentos do Fastgreedy (Figura C.4), apresenta uma estrutura mais homogênea.

A Figura G.5 e a Figura G.6 apresentam, respectivamente, os agrupamentos obtidos pelos métodos k – means e Fastgreedy. A base de dados com VM para o período de depois resultou em uma rede mais densa que convergiu para ambos os algoritmos com 5 agrupamentos. Portanto, neste caso a formação de sub-agrupamentos não foi tão evidente.

Figura G.5 – Agrupamentos k – means com VM de 2012-2016.

Figura G.6 – Agrupamentos Fastgreedy com VM de 2012-2016.

A Figura G.7 e a Figura G.8 apresenta, respectivamente, os grafos obtidos com VNF para o período de antes e depois. A Figura G.9 e G.10 ilustram, respectivamente, os agrupamentos obtidos com o k – means e o Fastgreedy com VNF para 2009-2012. Foram obtidos 4 e 3 agrupamentos, respectivamente, o que implica que a estrutura dos agrupamentos em grafos não tenha permitido a identificação da formação sub-agrupamentos.

Estas observações se estendem a análise dos grafos obtidos com VNF para 2012-2016 que convergem para um número próximo de agrupamentos. A Figura G.11 e a Figura G.12 ilustram, respectivamente, os 4 agrupamentos obtidos pelos métodos k – means e os 3 agrupamentos obtidos pelo Fastgreedy.

Figura G.7 – Grafo obtido com VNF de 2009-2012.

Figura G.8 – Grafo obtido com VNF de 2012-2016.

Figura G.9 - Agrupamentos k – means com VNF de 2009-2012.

Figura G.10 - Agrupamentos Fastgreedy com VNF de 2009-2012.

Figura G.11 - Agrupamentos k – means com VNF de 2012-2016.

Figura G.12 - Agrupamentos Fastgreedy com VNF de 2012-2016.

Os grafos obtidos com os bancos de dados de acidentes com VF para o cenário de antes e depois, são ilustrados respectivamente pela Figura G.13 e pela Figura G.14. E os agrupamentos para o cenário de antes e depois pelos métodos k – means e Fastgreedy são ilustrados, respectivamente, pelas Figuras G.15 e G.16 e pelas Figuras G.17 e G.18.

Figura G.13 – Grafo obtido com VF de 2009-2012.

Figura G.14 – Grafo obtido com VF de 2012-2016.

Figura G.15 - Agrupamentos k – means com VF de 2009-2012.

Figura G.16 - Agrupamentos Fastgreedy com VF de 2009-2012.

Figura G.17 - Agrupamentos k – means
com VF de 2012-2016.

Figura G.18 - Agrupamentos Fastgreedy
com VF de 2012-201.

Apêndice H – Regras de decisão obtidas com os agrupamentos derivados dos métodos de particionamento k – means e Fastgreedy

Tabela 18 - Regras de decisão obtidas a partir das técnicas de agrupamento de dados exploradas na pesquisa.

Base	Método	Agrup.	Regra	Nó	Conf.	Acerto (%)	Acurácia (%)
VM Antes	k-means	1	Acostamento [existe] + Sentido [Norte] + km [≤ 138] + Barreiras [Existe e Não Existe] + Largura da faixa Crítica [$\leq 3,52$ m]	15	0,98	50	96
		1	Acostamento [existe] + Sentido [Sul] + km [$\leq 133,900$] + Perfil da Pista [Declive]	12	0,96		
		2	Acostamento [existe] + Sentido [Norte] + km [≤ 138] + Barreiras [Existe e Não Existe] + Largura da Faixa crítica [$\leq 3,52$ m]	9	0,99		
		2	Acostamento [existe] + Sentido [Norte] + km [≤ 138] + Barreiras [Existe e Não existe] + Largura da faixa crítica [$> 3,52$ m]	16	0,77		
		2	Acostamento [existe] + Sentido [Norte] + km [> 138]	6	0,10		
		3	Acostamento [existe] + Sentido [Sul] + km [$> 133,900$] + Perfil da pista [Em nível ou Aclive] + Largura da Faixa Crítica [$> 2,52$]	20	0,10		
VM Antes	Fastgreedy	4	Acostamento [não existe]	2	0,10		
		1	Sentido da via [Sul] + Perfil da pista [Em nível ou Aclive] + Tipo de Acidente [colisão frontal, colisão transversal, colisão lateral, engavetamento, capotamento, atropelamento, tombamento, choque e outros]	8	0,10		
		1	Sentido da via [Sul] + Perfil da pista [Em nível] + Tipo de Acidente [Colisão Traseira] + km [$\leq 141,900$]	15	1,00		
		1	Sentido da via [Sul] + Perfil da pista [Em nível] + Tipo de Acidente [Colisão Traseira] + km [$> 141,900$]	16	0,88		
		2	Sentido da via [Norte] + Barreiras [Existe ou Misto] + Largura da Faixa Crítica [$> 2,96$ m] + Velocidade máxima [> 100 km/h]	19	1,00		
		3	Sentido da via [Sul] + Perfil da pista [Aclive ou Declive] + Largura da Faixa Crítica [$\leq 3,56$ m]	17	0,79		
		3	Sentido da via [Norte] + Barreiras [Existe ou Não Existe] + km [> 130]	12	0,80		
		3	Sentido da via [Norte] + Barreiras [Existe ou Misto] + Largura da Faixa Crítica [$\leq 2,96$ m]	13	0,90		

		3	Sentido da via [Norte] + Barreiras [Existe ou Misto] + Largura da Faixa Crítica [$> 2,96$ m] + Velocidade [100 km/h]	19	1,00		
VM Depois	k-means	1	km [> 133] + Barreiras [Não existe] + Sentido [Norte]	7	0,99	71	93
		2	km [> 133] + Barreiras [Existe ou Misto] + Largura do Acostamento [≤ 0 m] + Causa Provável [Motorista ou Veículo]	11	0,89		
		3	km [$\leq 132,800$]	3	1,00		
		4	km [> 133] + Barreiras [Não existe] + Sentido [Sul]	8	1,00		
		4	km [> 133] + Barreiras [Existe ou Misto] + Causa Provável [Via/meio ambiente e Outros] + Sentido [Sul]	16	1,00		
		4	km [> 133] + Barreiras [Existe ou Misto] + Largura do Acostamento [> 0 m] + Perfil da Pista [Aclive e Declive] + Largura da Faixa Crítica [$> 3,29$ m]	20	0,56		
		5	km [> 133] + Barreiras [Existe ou Misto] + Largura do Acostamento [> 0 m] + Perfil da Pista [Aclive e Declive] + km [$> 134,550$]	17	0,49		
VM Depois	Fastgreedy	1	km [$> 132,900$] + Barreiras [Não existe] + Sentido [Norte]	8	1,00	86	93
		2	km [$> 132,700$]	4	0,70		
		3	km [$\leq 132,700$]	3	1,00		
		4	km [$> 132,900$] + Barreiras [Não Existe] + Sentido [Norte]	7	0,99		
		4	km [$> 132,900$] + Barreiras [Existe ou Misto] + Largura do Acostamento [≤ 0] + Causa Provável [Via/Meio Ambiente ou Outros] + Sentido [Norte]	16	1,00		
		4	km [$> 132,900$] + Barreiras [Existe ou Misto] + Largura do Acostamento [> 0] + Perfil da Pista [Aclive ou Declive] + Largura da Faixa Crítica [$> 2,66$ m]	20	0,56		
		5	km [$> 132,900$] + Barreiras [Existe ou Misto] + Largura do Acostamento [> 0] + Perfil da Pista [Em nível] + km [$> 134,550$]	18	0,99		
VNF Antes	k-means	1	Sentido [Sul] + km [$\leq 141,100$]	3	0,84	100	89
		2	Sentido [Sul] + km [$> 141,100$]	4	0,89		
		3	Sentido [Norte] + Inclinação da pista [$> 2,02\%$]	5	0,97		
		4	Sentido [Norte] + Inclinação da pista [$\leq 2,02\%$]	6	0,87		
VNF Antes	Fastgreedy	1	Barreiras [Existe e Não existe] + Sentido [Sul] + VDM_Passeio [≤ 14.200]	11	1,00	100	94
		2	Barreiras [Não existe] + Largura do Acostamento [$> 2,67$ m]	4	0,96		
		3	Barreiras [Existe e Não existe] + Sentido [Norte] + Largura da Faixa Crítica [$\leq 3,45$ m]	10	1,00		

Base	Método	Agrup.	Regra	Nó	Conf.	Acerto (%)	Acurácia (%)
VNF Depois	k-means	1	km [> 133 km] + Causa provável [via/meio ambiente ou outros] + Sentido [Sul]	5	0,97	100	98
		2	km [<= 133]	1	0,96		
		3	km [> 133] + Causa Provável [Motorista ou veículo]	3	0,99		
		4	km [> 133] + Causa Provável [via/meio ambiente ou outros] + Sentido [Norte]	6	1,00		
VNF Depois	Fastgreedy	1	km [> 133] + Causa Provável [via/meio ambiente ou outros] + Sentido [Sul]	3	0,93	100	92
		2	km [<= 133]	1	0,96		
		3	km [> 133] + Causa Provável [Motorista ou veículo]	6	0,87		
		4	km [> 133] + Causa Provável [via/meio ambiente ou outros] + Sentido [Norte]	5	0,90		
VF Antes	k-means	1	Sentido [Sul]	1	0,72	33	72
		2	Sentido [Norte]	2	0,70		
		3	Sentido [Norte]	2	0,70		
VF Antes	Fastgreedy	1	km [<= 138]	1	0,89	100	83
		2	km [>138]	2	0,87		
		3	km [>138]	2	0,88		
VF Depois	k-means	1	km [<= 132,600]	1	0,86	100	86
		2	km [> 132,600]	2	0,86		
VF Depois	Fastgreedy	1	Causa provável [veículo ou via/meio ambiente ou outros] + km [<= 133,900]	3	0,75	100	84
		2	Causa provável [veículo ou via/meio ambiente ou outros] + km [> 133,900]	4	0,92		
		3	Causa provável [veículo ou via/meio ambiente ou outros] + km [> 133,900]	4	0,96		
		4	Causa provável [Motorista]	1	0,72		

Apêndice I – Árvores de decisão e matriz confusão obtidas a partir dos agrupamentos detectados com os métodos k – means e Fastgreedy

Neste Apêndice são apresentadas as árvores de decisão geradas a partir da classificação dos agrupamentos obtidos com os métodos de particionamento de observações explorados nas seções 4.2 e 4.3.

As regras de decisão foram extraídas a partir do algoritmo CART com as bases de dados balanceadas, uma vez que os classificadores são sensíveis ao desequilíbrio de classes. Esta característica inerente aos bancos de dados explorados em Segurança Viária conduz a obtenção de resultados subestimados, uma vez que os classificadores tendem a se ajustar em demasia as classes mais abundantes presentes nas bases de dados, apresentando uma baixa taxa de acerto na classificação das classes mais raras.

A Tabela I.1 apresenta a matriz de erro de classificação, também denominada de matriz confusão, para a base de acidentes com VM para o período de 2009-2012, com um total de 1049 observações atribuídas a cada agrupamento obtido com o k – means. Obteve-se uma acurácia geral de 96,4%, valor este dentro do que é recomendado pela literatura.

Tabela I.1 – Matriz confusão da base de dados com VM balanceada de 2009-2012 com os agrupamentos obtidos com o k – means.

Categoria	Classificação				Acerto (%)
	Agrup.1	Agrup.2	Agrup.3	Agrup.4	
Agrup.1	1031	1	17	0	98,3
Agrup.2	64	985	0	0	93,9
Agrup.3	71	0	978	0	93,2
Agrup.4	0	0	0	1049	100,0
Geral (%)	27,8	23,5	23,7	25,0	96,4

Na Tabela I.1 os valores presentes na diagonal principal correspondem as observações devidamente classificadas em cada agrupamento, ou seja, os verdadeiros positivos. Enquanto que os demais valores apresentam o número de observações atribuídas erroneamente a um determinado grupo (falso positivo). A Figura I.1 apresenta os resultados da árvore de decisão obtida a partir da base de dados balanceada com VM para o cenário de antes com o algoritmo de particionamento k – means.

Figura I.1 – Resultados da árvore de decisão com VM para o cenário de antes com os agrupamentos obtidos pelo k – means.

A Tabela I.2 apresenta a matriz confusão com VM para o período de 2009-2012, com um total de 1140 observações atribuídas a cada agrupamento obtido com o Fastgreedy. Verifica-se que a acurácia geral obtida foi de 95,6%. E a Figura I.2 apresenta a árvore de decisão obtida com os agrupamentos considerando VM e o algoritmo Fastgreedy.

Tabela I.2 – Matriz confusão da base de dados com VM balanceada de 2009-2012 com os agrupamentos obtidos com o Fastgreedy.

Categoria	Classificação			
	Agrup.1	Agrup.2	Agrup.3	Acerto (%)
Agrup.1	1077	0	63	94,5
Agrup.2	0	1111	29	97,5
Agrup.3	58	0	1082	94,9
Geral (%)	33,2	32,5	34,3	95,6

A classificação dos agrupamentos obtidos com a base de dados de VM para o período de depois com o k – means apresentou acurácia da ordem de 92,7%, conforme se verifica na Tabela I.3, com o total de 655 observações atribuídas a cada agrupamento. A Figura I.3 apresenta a árvore de decisão a partir dos resultados obtidos.

Tabela I.3 – Matriz confusão da base de dados com VM balanceada de 2012-2016 com os agrupamentos obtidos com o k – means.

Categoria	Classificação					Correto (%)
	Agrup.1	Agrup.2	Agrup.3	Agrup.4	Agrup.5	
Agrup.1	655	0	0	0	0	100,0
Agrup. 2	0	627	17	10	1	95,7
Agrup.3	0	0	604	51	0	92,2
Agrup.4	0	10	0	643	2	98,2
Agrup. 5	0	120	0	28	507	77,4
Geral (%)	20,0	23,1	19,0	22,4	15,6	92,7

A matriz confusão obtida a partir da classificação dos agrupamentos com o *Fast Greedy* para VM no cenário de depois é apresentada na Tabela I.4. Verifica-se que a que a matriz confusão apresenta a mesma taxa de acerto do k – means (Tabela I.3), bem como a mesma acurácia geral de 92,7%.

As estruturas das árvores de decisão obtidas com o k – means e o *Fast Greedy* são bem semelhantes. Porém, as regras de decisão extraídas com o *Fast Greedy* foram diferentes e apresentam, em geral, a maior probabilidade de acerto (Apêndice H). A Figura I.4 apresenta a árvore de decisão obtida a partir destes resultados.

Tabela I.4 – Matriz confusão da base de dados com VM balanceada de 2012-2016 com os agrupamentos obtidos com o Fastgreedy.

Categoria	Classificação					Correto (%)
	Agrup.1	Agrup.2	Agrup.3	Agrup.4	Agrup.5	
Agrup.1	655	0	0	0	0	100,0%
Agrup.2	0	627	17	10	1	95,7%
Agrup.3	0	0	604	51	0	92,2%
Agrup.4	0	10	0	643	2	98,2%
Agrup.5	0	120	0	28	507	77,4%
Geral (%)	20,0	23,1	19,0	22,4	15,6	92,7%

Em relação a base de dados com VNF para o período de antes obteve-se a acurácia geral de 89,10% a partir dos agrupamentos obtidos com o algoritmo k – means, conforme se verifica na Tabela I.5.

Tabela I.5 – Matriz confusão da base de dados com VNF balanceada de 2009-2012 com os agrupamentos obtidos com o k – means.

Categoria	Classificação				Correto (%)
	Agrup.1	Agrup.2	Agrup.3	Agrup.4	
Agrup.1	211	23	0	0	90,2%
Agrup.2	40	194	0	0	82,9%
Agrup.3	0	0	228	6	97,4%
Agrup.4	0	0	33	201	85,9%
Geral (%)	26,8%	23,2%	27,9%	22,1%	89,1%

A árvore de decisão obtida para VNF no cenário de antes com o algoritmo k – means é apresentada na Figura I.5.

A extração de regras de decisão com VNF a partir dos agrupamentos obtidos com o *Fast Greedy* para o cenário de antes apresentou acurácia geral de 93,8%, conforme ilustra a Tabela I.6.

Tabela I.6 – Matriz confusão da base de dados com VNF balanceada de 2009-2012 com os agrupamentos obtidos com o Fastgreedy.

Categoria	Classificação			Correto (%)
	Agrup.1	Agrup.2	Agrup.3	
Agrup.1	195	15	16	86,3%
Agrup.2	4	222	0	98,2%
Agrup.3	0	7	219	96,9%
Geral (%)	29,4%	36,0%	34,7%	93,8%

A árvore de decisão obtida com a base VNF e a partir dos agrupamentos do *Fast Greedy* para o cenário de antes é apresentada na Figura I.6.

Figura I.2 – Resultados da árvore de decisão com VM para o cenário de antes com os agrupamentos obtidos pelo Fastgreedy.

Figura I.3 – Resultados da árvore de decisão com VM para o cenário de depois com os agrupamentos obtidos pelo k – means.

Figura I.4 – Resultados da árvore de decisão com VM para o cenário de depois com os agrupamentos obtidos pelo Fastgreedy.

Figura I.5 – Resultados da árvore de decisão com VNF para o cenário de antes com os agrupamentos obtidos pelo k – means.

Figura I.6 – Resultados da árvore de decisão com VNF para o cenário de antes com os agrupamentos obtidos pelo Fastgreedy.

Para os agrupamentos obtidos com o k – means com VNF para o cenário de depois resultou em uma acurácia geral da ordem de 98,3%, conforme ilustra a Tabela I.7.

Tabela I.7 – Matriz confusão da base de dados com VNF balanceada de 2012-2016 com os agrupamentos obtidos com o k – means.

Categoria	Classificação				Correto (%)
	Agrup.1	Agrup.2	Agrup.3	Agrup.4	
Agrup.1	204	0	0	0	100.0%
Agrup.2	5	197	2	0	96.6%
Agrup.3	0	2	202	0	99.0%
Agrup.4	0	5	0	199	97.5%
Geral (%)	25.6%	25.0%	25.0%	24.4%	98.3%

Figura I.7 – Resultados da árvore de decisão com VNF para o cenário de depois com os agrupamentos obtidos pelo k – means.

A acurácia na classificação obtida a partir dos agrupamentos com VNF a partir do *Fast Greedy* foi de 92,0%, conforme se verifica na Tabela I.8

Tabela I.8 – Matriz confusão da base de dados com VNF balanceada de 2012-2016 com os agrupamentos obtidos com o Fastgreedy.

Classificação					
Categoria	Agrup.1	Agrup.2	Agrup.3	Agrup. 4	Correto (%)
Agrup.1	157	0	20	14	82,2%
Agrup.2	0	184	5	2	96,3%
Agrup.3	0	3	186	2	97,4%
Agrup.4	11	3	1	176	92,1%
Geral (%)	22,0%	24,9%	27,7%	25,4%	92,0%

Figura I.8 – Resultados da árvore de decisão com VNF para o cenário de depois com os agrupamentos obtidos pelo Fastgreedy.

A base de dados com VF resultou em uma acurácia de 71,8% com a classificação dos agrupamentos obtidos com o método k – means, conforme demonstra a Tabela I.9 e a árvore de decisão na Figura I.9.

Tabela I.9 – Matriz confusão da base de dados com VF de 2009-2012 com os agrupamentos obtidos com o k – means.

Classificação				
Categoria	Agrup.1	Agrup.2	Agrup.3	Correto (%)
Agrup.1	16	0	0	72,4
Agrup.2	0	8	8	70,0
Agrup.3	0	12	4	70,0
Geral (%)	34,3	33,7	32,0	71,8

Figura I.9 – Resultados da árvore de decisão com VF para o cenário de antes com os agrupamentos obtidos pelo k – means.

Para o cenário de antes usando os agrupamentos obtidos com o *Fast Greedy* obteve-se uma acurácia geral de 83,3% e a matriz confusão apresentada na Tabela I.10. A árvore de decisão com os resultados obtidos é apresentada na Figura I.10.

Tabela I.10 – Matriz confusão da base de dados com VF de 2009-2012 com os agrupamentos obtidos com o Fastgreedy.

Classificação				
Categoria	Agrupamento 1	Agrupamento 2	Agrupamento 3	Correto (%)
Agrup.1	13	0	2	93,5
Agrup.2	3	12	0	81,2
Agrup.3	2	11	2	75,2
Geral (%)	34,7	36,0	29,4	83,3

Figura I.10 – Resultados da árvore de decisão com VF para o cenário de antes com os agrupamentos obtidos pelo Fastgreedy.

No cenário de depois com VF obteve-se a acurácia geral de 86,0% com o k – means, conforme verifica-se na Tabela I.11. A árvore de decisão obtida é apresentada na Figura I.11.

Tabela I.11 – Matriz confusão da base de dados com VF de 2012-2016 com os agrupamentos obtidos com o k – means.

Categoria	Classificação		
	Agrup.1	Agrup.2	Correto (%)
Agrup.1	15	4	86,0
Agrup.2	2	17	86,0
Geral (%)	42,1	57,9	86,0

Figura I.11 – Resultados da árvore de decisão com VF para o cenário de depois com os agrupamentos obtidos pelo k – means.

O *Fast Greedy* resultou com a base de VF depois uma acurácia geral da ordem de 84,2%, sendo a matriz confusão apresentada na Tabela I.12. A árvore de decisão com os resultados obtidos é apresentada na Figura I.12

Tabela I.12 – Matriz confusão da base de dados com VF de 2012-2016 com os agrupamentos obtidos com o Fastgreedy.

Categoria	Classificação				Correto (%)
	Agrup.1	Agrup.2	Agrup.3	Agrup. 4	
Agrup.1	9	1	0	0	75,7
Agrup.2	0	10	0	0	92,0
Agrup.3	1	0	7	0	72,6
Agrup.4	0	0	0	10	96,5
Geral (%)	22,0%	27,7%	24,9%	25,4%	84,2

Figura I.12 – Resultados da árvore de decisão com VF para o cenário de depois com os agrupamentos obtidos pelo Fastgreedy.

Apêndice J – Fatores discriminantes na formação dos agrupamentos obtidos pelas técnicas k-means e Fastgreedy

Neste Apêndice são apresentadas as variáveis relacionadas na formação de cada agrupamento, bem como sua respectiva taxa de probabilidade. Em negrito são destacadas as que apresentam as maiores probabilidades, bem como as variáveis mais destoantes, ou seja, que embora nem sempre apresentem alta probabilidade de ocorrência (acima de 70%) quando analisada em conjunto com as demais variáveis permitem a caracterização ou discriminação geral e eficiente dos agrupamentos obtidos.

A Tabela J.1 apresenta as características dominantes na formação de cada agrupamento, a partir das bases de dados desbalanceadas com VM nos cenários antes e depois, e suas respectivas probabilidades que podem ser utilizadas para caracterizar cada agrupamento visando reduzir a heterogeneidade.

Tabela J.1 – Fatores discriminantes e suas probabilidades em cada agrupamento com VM no cenário antes e depois.

Variável: valor do atributo	Agrupamentos							
	DM antes fastgreedy			DM depois fastgreedy				
	Agrup.1 (%)	Agrup.2 (%)	Agrup.3 (%)	Agrup.1 (%)	Agrup.2 (%)	Agrup.3 (%)	Agrup.4 (%)	Agrup.5 (%)
Tipo de acidente: colisão traseira	35	34	33	35	35	30	28	30
Tipo de acidente: choque	21	27	24	11	26	30	24	34
Tipo de acidente: colisão lateral	11	10	17	18	9	11	19	9
Tipo de acidente: engavetamento	10	9	11	26	7	11	15	15
Causa provável: motorista	100	63	68	52	100	40	35	46
Causa provável: outros	0	37	32	44	0	60	65	47
Sentido da via: Norte	0	100	0	100	31	46	61	100
Sentido da via: Sul	100	0	100	0	69	54	39	0
Traçado: Reto	95	96	91	100	72	93	91	96
Número de faixas: 3	100	100	100	100	100	85	100	100
Largura da faixa crítica: 2,80 m a 4,15 m	100	100	100	86	71	73	76	75
Local do acidente: Faixa de rolamento	100	76	78	75	87	95	74	71
Canteiro central: Existe	100	100	100	100	100	100	100	100
Acostamento: Existe	100	100	100	100	100	100	100	100
Radar: Não existe	100	100	100	100	0	100	0	0
Radar: Existe	0	0	0	0	100	0	100	100
Período: manhã	50	50	39	52	77	40	30	46
Período: tarde	50	50	41	4	0	9	0	7

Variável: valor do atributo	Agrupamentos							
	DM antes fastgreedy			DM depois fastgreedy				
	Agrup. 1 (%)	Agrup. 2 (%)	Agrup. 3 (%)	Agrup. 1 (%)	Agrup. 2 (%)	Agrup. 3 (%)	Agrup. 4 (%)	Agrup. 5 (%)
Período: noite	0	0	21	44	23	51	70	47
Condição de visibilidade: boa	55	55	60	81	69	58	64	58
Condição de visibilidade: parcial	45	45	40	19	30	41	36	41
Condição do tempo: bom	84	82	83	91	90	88	89	84
Condição da pista: seca	86	82	83	91	90	88	89	84
Trevo: Existe	50	0	49	100	100	7	100	100
Trevo: Não existe	50	100	51	0	0	93	0	0
Dispositivo de acesso: entrada	25	37	27	69	20	25	25	25
Dispositivo de acesso: saída	24	16	28	16	32	27	27	13
Dispositivo de acesso: entrada/saída	51	20	45	15	33	48	24	62
Dispositivo de acesso: não existe	0	27	0	0	0	0	23	0
Velocidade máxima: 100 km/h	98	80	71	75	100	69	88	100
Velocidade máxima: 110 km/h	2	20	29	25	0	31	12	0
Perfil: em nível e Inclinação: 0% a 2%	81	68	50	70	48	39	76	93
Perfil: aclive 4,34 % A 4,58%	10	32	50	30	0	14	16	7
Perfil: declive	9	0	0	0	41	47	8	0
Obras: Não existe	100	100	100	100	100	98	97	96
SH: existe	100	100	100	100	100	98	97	96
SV: existe	100	100	100	100	100	98	97	96
Barreiras: Existe	0	42	22	0	47	22	39	14
Barreiras: Misto	80	24	46	0	53	76	17	86
Barreiras: Não existe	20	34	33	86	0	2	44	0
Larg. do acostamento: 2,60 m a 3,71 m	72	78	74	86	77	82	70	73
Largura do canteiro central: 10,64 m a 13,52 m	100	100	100	86	61	88	100	30
Iluminação: não existe	100	100	100	100	100	100	100	100

A Tabela J.2 apresenta as características dominantes na formação de cada agrupamento, a partir das bases de dados desbalanceada com VNF nos cenários antes e depois.

Tabela J.2 - Fatores discriminantes e suas probabilidades em cada agrupamento com VNF no cenário antes e depois.

Variável: valor do atributo	Agrupamentos						
	NF antes fastgreedy			NF depois k-means			
	Agru.1 (%)	Agrup. 2 (%)	Agrup. 3 (%)	Agrup. 1 (%)	Agrup. 2 (%)	Agrup. 3 (%)	Agrup. 4 (%)
Tipo de acidente: colisão traseira	26	35	26	22	26	27	19
Tipo de acidente: choque	20	22	24	19	21	18	27
Tipo de acidente: colisão lateral	10	8	2	27	13	20	19
Tipo de acidente: outros	8	5	3	22	12	19	14
Causa provável: motorista	75	64	66	3	31	92	5
Causa provável: outros	12	15	14	97	61	8	95
Sentido da via: Norte	27	26	100	3	48	52	100
Sentido da via: Sul	73	74	0	97	52	48	0
Traçado: Reto	89	98	98	94	94	89	86
Número de faixas: 3	99	99	99	96	81	98	100
Largura da faixa crítica: 2,78 m a 3,65 m	43	79	79	27	40	11	57
Local do acidente: Faixa de rolamento	66	78	73	76	79	81	78
Canteiro central: Existe	95	100	100	96	100	100	100
Acostamento: Existe	89	100	100	100	100	100	100
Radar: Não existe	100	100	100	97	100	100	100
Radar: Existe	0	0	0	3	0	0	0
Período: manhã	39	52	36	2	30	20	40
Período: tarde	41	36	49	1	9	15	22
Período: noite	20	12	18	97	61	65	38
Condição de visibilidade: boa	52	72	62	70	56	67	60
Condição de visibilidade: parcial	47	28	40	30	43	27	40
Condição do tempo: bom	83	92	90	94	98	91	90
Condição da pista: seca	81	93	90	94	88	91	90
Trevo: Existe	44	0	0	93	100	100	100
Trevo: Não existe	56	100	100	7	0	0	0
Dispositivo de acesso: entrada	23	1	1	43	50	33	31
Dispositivo de acesso: saída	10	25	25	25	4	3	19
Dispositivo de acesso: entrada/saída	67	41	41	32	46	64	50
Dispositivo de acesso: não existe	0	33	33	0	0	0	0
Velocidade máxima: 100 km/h	100	98	98	97	65	60	98
Velocidade máxima: 110 km/h	0	2	2	3	35	40	2
Perfil: em nível e Inclinação: 0% a 2%	28	91	91	79	80	77	98
Perfil: acline 4,34 % A 4,58%	34	9	9	11	10	23	1
Perfil: declive	38	0	0	11	10	0	2
Obras: Não existe	96	100	100	96	98	98	99
Sinalização Horizontal: existe	96	100	100	96	98	98	99
Sinalização Vertical: existe	96	100	100	96	98	98	99
Barreiras: Existe	20	10	82	35	29	25	53
Barreiras: Misto	80	8	8	27	71	43	47
Barreiras: Não existe	0	82	10	38	0	32	0

Variável: valor do atributo	Agrupamentos						
	NF antes fastgreedy			NF depois k-means			
	Agrup.1 (%)	Agrup.2 (%)	Agrup.3 (%)	Agrup.1 (%)	Agrup.2 (%)	Agrup.3 (%)	Agrup.4 (%)
Largura do acostamento: 2,79 m a 3,10 m	37	81	81	10	35	20	61
Largura do canteiro central: 10,64 m a 13,7 m	80	58	58	46	46	30	26
Iluminação: não existe	79	100	100	100	100	100	100

A Tabela J.3 apresenta as características dominantes na formação de cada agrupamento, a partir das bases de dados desbalanceada com VF nos cenários antes e depois.

Tabela J.3 - Fatores discriminantes e suas probabilidades em cada agrupamento com VF no cenário antes e depois.

Variável: valor do atributo	Agrupamentos				
	VF antes fastgreedy			VF depois k-means	
	Agrup.1 (%)	Agrup.2 (%)	Agrup.3 (%)	Agrup.1 (%)	Agrup.2 (%)
Tipo de acidente: atropelamento	67	64	22	70	42
Tipo de acidente: tombamento	13	11	15	13	11
Tipo de acidente: colisão frontal	12	9	33	12	8
Tipo de acidente: capotamento	8	15	12	20	5
Tipo de acidente: colisão traseira	2	5	6	12	32
Causa provável: motorista	33	29	22	10	32
Causa provável: outros	67	71	78	90	68
Sentido da via: Norte	67	36	100	62	68
Sentido da via: Sul	33	64	0	38	37
Traçado: Reto	100	100	100	100	74
Traçado: curva	0	0	0	0	26
Número de faixas: 2	0	0	0	30	0
Número de faixas: 3	100	100	100	60	100
Número de faixas: 4	0	0	0	10	0
Largura da faixa crítica: 3,00 m a 3,65 m	65	78	77	20	70
Local do acidente: alça de saída	67	0	2	0	14
Local do acidente: acostamento	27	0	22	10	2
Local do acidente: faixa de rolamento	10	93	78	90	79
Canteiro central: Existe	100	100	100	100	100
Acostamento: Existe	87	100	100	100	100

Variável: valor do atributo	Agrupamentos				
	VF antes fastgreedy			VF depois k-means	
	Agrup.1 (%)	Agrup.2 (%)	Agrup.3 (%)	Agrup.1 (%)	Agrup.2 (%)
Acostamento: Não Existe	13	0	0	0	0
Radar: Não existe	100	100	100	90	100
Radar: Existe	0	0	0	10	0
Período: manhã	33	30	78	10	32
Período: tarde	13	0	9	0	0
Período: noite	53	70	13	90	68
Condição de visibilidade: boa	20	29	33	20	37
Condição de visibilidade: parcial	73	71	56	80	63
Condição de visibilidade: ruim	7	0	11	0	0
Condição do tempo: bom	87	93	89	100	95
Condição do tempo: chuvoso	13	25	11	0	5
Condição da pista: seca	87	93	89	100	95
Condição da pista: molhada	13	36	11	0	5
Trevo: Existe	73	14	100	100	100
Trevo: Não existe	27	86	0	0	0
Dispositivo de acesso: entrada	13	29	11	0	37
Dispositivo de acesso: saída	27	36	33	0	0
Dispositivo de acesso: entrada/saída	40	36	56	34	47
Dispositivo de acesso: não existe	20	0	0	0	16
Velocidade máxima: 90 km/h	0	0	33	0	0
Velocidade máxima: 100 km/h	73	100	56	50	89
Velocidade máxima: 110 km/h	27	0	11	50	11
Perfil: em nível e Inclinação: 0% a 2%	53	93	80	20	68
Perfil: aclive 4,34 % A 4,58%	0	7	14	70	10
Perfil: declive	47	0	6	10	22
Obras: Não existe	100	100	100	100	100
Sinalização Horizontal: existe	100	100	100	100	100
Sinalização Vertical: existe	100	100	100	100	100
Barreiras: Existe	27	21	12	60	53
Barreiras: Misto	67	79	80	40	37
Barreiras: Não existe	6	0	8	0	11
Largura do acostamento: 3,00 m a 3,50 m	38	78	100	100	100
Largura do canteiro central: 9,04 m a 13,17 m	67	100	100	100	100

Agrupamentos					
Variável: valor do atributo	VF antes fastgreedy			VF depois k-means	
	Agrup.1 (%)	Agrup.2 (%)	Agrup.3 (%)	Agrup.1 (%)	Agrup.2 (%)
Iluminação: não existe	100	100	67	100	100
Iluminação: Existe	0	0	33	0	0

Apêndice K – Variáveis contribuintes a ocorrência dos acidentes no período “antes” e “depois” da concessão da rodovia SP-065 obtidas por de regras de classificação utilizando a abordagem de BNN a partir de bases de dados de acidentes rodoviários (em CD-ROM).