UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

DIEGO VIEIRA NEVES

Uso de aprendizado supervisionado para análise de confiabilidade de dados
de *crowdsourcing* sobre posicionamento de ônibus

São Paulo

2018

DIEGO VIEIRA NEVES

**Uso de aprendizado supervisionado para análise de confiabilidade de dados
de *crowdsourcing* sobre posicionamento de ônibus**

Dissertação apresentada à Escola de Artes,
Ciências e Humanidades da Universidade de
São Paulo para obtenção do título de Mestre
em Ciências pelo Programa de Pós-graduação
em Sistemas de Informação.

Área de concentração: Metodologia e
Técnicas da Computação

Versão corrigida contendo as alterações
solicitadas pela comissão julgadora em 16 de
outubro de 2018. A versão original encontra-
se em acervo reservado na Biblioteca da
EACH-USP e na Biblioteca Digital de Teses
e Dissertações da USP (BDTD), de acordo
com a Resolução CoPGr 6018, de 13 de
outubro de 2011.

Orientador: Prof. Dr. Daniel Cordeiro

São Paulo

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

CRB 8 - 4936

Neves, Diego Vieira

Uso do aprendizado supervisionado para análise de confiabilidade de dados de *crowdsourcing* sobre posicionamento de ônibus / Diego Vieira Neves ; orientador, Daniel Cordeiro. – 2018.

87 f. : il

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.

Versão corrigida

1. Aprendizado computacional. 2. Sistemas de transportes. 3. Sistemas de informação. 4 Transporte público. 5. Ônibus. I. Cordeiro, Daniel de Angelis, orient. II. Título

CDD 22.ed.– 006.31

Dissertação de autoria de Diego Vieira Neves, sob o título “**Uso de aprendizado supervisionado para análise de confiabilidade de dados de *crowdsourcing* sobre posicionamento de ônibus**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 16 de outubro de 2018 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Daniel de Angelis Cordeiro

Instituição: Universidade de São Paulo

Presidente

Prof. Dr. Renato Porfirio Ishii

Instituição: Universidade Federal de Mato Grosso do Sul

Prof. Dr. Fabio Moreira Costa

Instituição: Universidade Federal de Goiás

Profa. Dra. Ana Amélia Benedito Silva

Instituição: Universidade de São Paulo

Agradecimentos

A Deus, pelas oportunidades concedidas, pelo discernimento em reconhecê-las e pelas pessoas extraordinárias que colocou em meu caminho que propiciaram, de forma direta ou indireta, a conclusão de mais uma etapa da minha vida.

Ao meu orientador Prof. Dr. Daniel de Angelis Cordeiro, pela orientação, incentivo e ensinamentos durante as atividades de mestrado, pela ajuda nesta dissertação, por proporcionar as condições necessárias para o desenvolvimento deste trabalho, pela confiança depositada e principalmente pelos momentos de positivismo e amizade.

A todos os professores, que contribuíram com a minha trajetória acadêmica, especialmente a Prof. Dr.^a Soraia Silva Prietch minha ex-orientadora do curso de graduação. Agradeço pela amizade e pelos ensinamentos e incentivo, e pelo exemplo como educadora que fez toda a diferença e conduziram meus passos até aqui.

Aos meus familiares por todo amor e carinho, e por entenderem os momentos de ausência. Sou grato, especialmente a meu pai Antônio Neves e meus avós Maria Neves e Dorecil Neves pela educação que me foi dada e por ter me ensinado valores que carrego comigo em todos os momentos. Dedico um agradecimento especial aos meus primos Kelly Neves e Lucas Neves, e minha tia Sônia Neves grandes parceiros e incentivadores.

A todos os meus amigos especialmente Igor de Oliveira, Adriano Santos e sua esposa Mariana Sales pela longa amizade, pelos inúmeros conselhos, puxões de orelha e por nunca negarem apoio e incentivo nos momentos difíceis. Sou grato a meus amigos e companheiros de mestrado Jonas Targino, Fernando Costa, Lígia Moreno, Williane Rodrigues, Felipe Cordeiro e Charles Mendes pelos momentos compartilhados, pelas palavras de incentivo e por não me deixaram ser vencido pelo cansaço. Agradeço ainda meus amigos e colegas de trabalho Karina Vaz, Mariana Rufino, Matheus Cañon, Leandro Corso, Paulo Abreu, Hugo Torralbo, Adriano Paixão, Edilson Albuquerque, Eduardo Boscovich e Regis Diniz por todo apoio, paciência e companheirismo.

Todos vocês foram fundamentais para concretização desse sonho, por isso merecem o meu eterno agradecimento.

Por fim, agradeço a empresa Scipopulis pelos dados fornecidos e o grupo de pesquisa INCT da Internet do Futuro para Cidades Inteligentes financiado pelo CNPq, proc. 465446/2014-0, CAPES, proc. 88887.136422/2017-00 e FAPESP, proc. 2014/50937-1.

Resumo

NEVES, Diego Vieira. **Uso de aprendizado supervisionado para análise de confiabilidade de dados de *crowdsourcing* sobre posicionamento de ônibus.** 2018. 87 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

Pesquisadores de diversas áreas estão estudando o desenvolvimento do que chamamos de Cidades Inteligentes: a integração de Sistemas de Informação e Comunicação com tecnologias de Internet das Coisas para utilizar os recursos de uma cidade de forma mais inteligente. Um dos principais objetivos das cidades inteligentes é solucionar os problemas relacionados à mobilidade urbana, que afeta significativamente a qualidade de vida da população. Um problema observável nas grandes metrópoles é a qualidade dos seus serviços de transporte público, especialmente quando nos referimos ao modal ônibus. A falta de informações confiáveis, associada à baixa qualidade dos serviços de transporte coletivo disponibilizados, leva o usuário a não optar pela utilização desse recurso, o que agrava problemas urbanos sociais e ambientais. Para reverter esse cenário, as iniciativas em cidades inteligentes propõem o uso de Sistemas de Transportes Inteligentes que podem utilizar diversos sensores e equipamentos para coletar diferentes tipos de dados referente aos serviços de transporte público. A captura e processamento desses dados permite, em tese, permite que o cidadão possa utilizar o transporte público com confiabilidade e previsibilidade. Contudo, esses dados podem ser insuficientes ou de baixa qualidade para uso em tempo real. Neste trabalho de mestrado investigamos o uso de dados obtidos via colaboração coletiva (*crowdsourcing*) como complemento dessas informações. Para mitigar as incertezas introduzidas pelo uso de *crowdsourcing*, este trabalho propõe a utilização de técnicas de aprendizado de máquina para criação de métodos de análise de confiabilidade dos dados coletados para o sistema de transporte público (por ônibus) do município de São Paulo. Para mitigar as incertezas introduzidas pelo uso de *crowdsourcing*, este trabalho propõe e compara o uso de diferentes técnicas de aprendizado de máquina para criar um modelo de análise de confiabilidade para os dados coletados, especializado no sistema de transporte coletivo (por ônibus) da cidade de São Paulo. Os resultados demonstram, que os algoritmos de Árvore de Decisão e Gaussian Naive Bayes foram mais eficazes e eficientes na realização da atividade de classificação dos dados obtidos com *crowdsourcing*. O algoritmo de Árvore de Decisão, apresentou os melhores indicadores de desempenho em termos de acurácia (94,34%) e F-score (99%), e o segundo melhor tempo de execução (0,023074 segundo). Já o algoritmo de Gaussian Naive Bayes foi o mais eficiente, com tempo médio de execução de 0,003182 segundos e foi o quarto melhor resultado em termos de acurácia (98,18%) e F-score (97%).

Palavras-chaves: *Crowdsourcing*; Aprendizado de máquina; Sistemas inteligentes de transporte; Cidades inteligentes.

Abstract

NEVES, Diego Vieira. **Use of supervised learning to analyze reliability of crowdsourcing bus location data**. 2018. 87 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

Researchers from different areas are studying the development of what we call Smart Cities: integrating Information and Communication Systems with Internet of Things to use city resources more intelligently. A major objective of smart cities is to solve problems related to urban mobility that significantly affects the quality of life of the population. An observable problem in big cities is the quality of their public transport services, specifically when we refer to the bus modal. The lack of reliable information, associated with the poor quality of public transport services, encouraging the user to look for alternatives, which aggravates urban social and environmental problems. To reverse this scenario, smart cities initiatives propose the use Intelligent Transport Systems, that can use various sensors and equipment to collect several types of data on public transport services. The capture and processing of these data allows, in theory, citizens to use the public transport with reliability and predictability. However, this data can be insufficient or of poor quality for usage in real-time. This master's work investigates the use of crowdsourcing data as a complement to this information. To mitigate the uncertainties introduced by the use of crowdsourcing, this research proposes and compares the use of different machine learning techniques to create a reliability analysis model for the data collected that is specialized for use on public transport system (bus) in the city of São Paulo. The results show that the Decision Tree and Gaussian Naive Bayes algorithms are more effective and efficient in performing the classification activity of the data obtained with crowdsourcing. The Decision Tree algorithm presented the best performance indicators in terms of accuracy (94.34%) and F-score (99%), and the second best execution time (0.023074 seconds). The Gaussian Naive Bayes algorithm was the most efficient, with an average execution time of 0.003182 seconds and was the forth best result in terms of accuracy (98.18%) and F-score (97%).

Keywords: Crowdsourcing; Machine learning; Intelligent transport systems; Smart cities.

Lista de figuras

Figura 1 – Quantidade de ônibus e linhas em algumas cidades do mundo.	20
Figura 2 – Arquitetura AVL (<i>Automatic Vehicle Location</i>).	22
Figura 3 – Configurações do aplicativo Coletivo.	28
Figura 4 – Representação do processo de classificação.	31
Figura 5 – Processo de classificação e sua utilização na predição de novos exemplos.	32
Figura 6 – Árvore de Decisão simples para análise de confiabilidade de dados sobre a localização dos ônibus.	35
Figura 7 – Exemplo de matriz de confusão.	39
Figura 8 – Exemplo de um gráfico de Curva ROC	41
Figura 9 – Exemplo de situação onde o dado informado pelo usuário não é confiável.	49
Figura 10 – Exemplo de situação onde o dado informado pelo usuário é confiável.	50
Figura 11 – Mapa de distribuição de localizações das contribuições coletivas.	53
Figura 12 – Coleta de dados de posicionamento dos ônibus.	53
Figura 13 – Etapas do processo de análise e tratamento dos dados.	54
Figura 14 – Divisão do conjunto de dados utilizando <i>K-fold cross-validation</i>	59
Figura 15 – Indicadores de desempenho do algoritmo de regressão logística.	63
Figura 16 – Indicadores de desempenho do modelo de <i>K-Nearest Neighbour</i>	64
Figura 17 – Indicadores de desempenho do modelo de Árvore de Decisão.	65
Figura 18 – Indicadores de desempenho do modelo de <i>Gaussian Naive Bayes</i>	67
Figura 19 – Indicadores de desempenho do modelo de Máquinas de Vetores Suporte.	68
Figura 20 – Comparação dos indicadores de desempenho dos classificadores.	70

Lista de quadros

Quadro 1 – Classificação dos modos de transporte urbano de passageiros.	19
Quadro 2 – Funções dos Sistemas de Informação ao Usuário.	23
Quadro 3 – Classificação de desempenho do modelo dado pela AUC.	42

Lista de tabelas

Tabela 1 – Exemplo de um amostra dos dados de <i>crowdsourcing</i>	31
Tabela 2 – Estudos que utilizaram técnicas de <i>crowdsourcing</i>	43
Tabela 3 – Descrição do conjunto de dados da SPTrans.	51
Tabela 4 – Descrição do conjunto de dados de <i>crowdsourcing</i>	52
Tabela 5 – Exemplos de registros dos dados de <i>crowdsourcing</i> analisados.	58
Tabela 6 – Conjunto de parâmetros utilizados.	61
Tabela 7 – Indicadores de desempenho apresentado pelos classificadores.	69
Tabela 8 – Execuções do modelo utilizando regressão logística.	83
Tabela 9 – Execuções do modelo utilizando <i>K-Nearest Neighbour</i>	84
Tabela 10 – Execuções do modelo utilizando Árvore de Decisão.	85
Tabela 11 – Execuções do modelo utilizando <i>Gaussian Naive Bayes</i>	86
Tabela 12 – Execuções do modelo utilizando Máquinas de Vetores Suporte.	87

Lista de abreviaturas e siglas

ANTP	Associação Nacional de Transportes Públicos
APTS	<i>Advanced Public Transportations</i>
AUC	Area Under the Curve
AVL	<i>Automatic Vehicle Location</i>
CART	<i>Classification on Regression Trees</i>
CV	<i>Cross-Validation</i>
GPS	<i>Global Positioning System</i>
GTFS	<i>General Transit Feed Specification</i>
ID3	<i>Iterative Dichotomiser 3</i>
ITS	<i>Intelligent Transport System</i>
KNN	<i>K-Nearest Neighbour</i>
MCC	<i>Matthews correlation coefficient</i>
ML	<i>Machine Learning</i>
ROC	<i>Receiver Operating Characteristics Curve</i>
SC	<i>Smart Cities</i>
SIM	Sistema Integrado de Monitoramento
SIU	Sistemas de Informação ao Usuário
SVM	<i>Support Vector Machine</i>
TAE	Teoria de Aprendizado Estatístico
TIC	Tecnologias da Informação e Comunicação

Sumário

1	Introdução	14
1.1	<i>Motivação e justificativa</i>	15
1.2	<i>Objetivo e hipótese</i>	15
1.3	<i>Estrutura do documento</i>	16
2	Cidades inteligentes, mobilidade urbana e suas tecnologias . .	18
2.1	<i>Cidades inteligentes</i>	18
2.2	<i>Conceitos sobre transporte urbano</i>	19
2.3	<i>Sistemas Inteligentes de Transporte</i>	21
2.4	<i>Sistemas de informação ao usuário</i>	22
3	Conceitos e tecnologias relacionados a crowdsourcing	24
3.1	<i>Crowdsourcing</i>	24
3.2	<i>Crowdsourcing em sistemas de informação ao usuário</i>	25
3.2.1	<i>Moovit</i>	26
3.2.2	<i>CityMapper</i>	26
3.2.3	<i>OneBusAway</i>	27
3.3	<i>Descrição do aplicativo Coletivo</i>	27
4	Conceitos e técnicas relacionadas a Aprendizado de Máquina	29
4.1	<i>Aprendizado de Máquina</i>	29
4.2	<i>Aprendizado supervisionado</i>	30
4.2.1	<i>Regressão Logística</i>	32
4.2.2	<i>K-Nearest Neighbour</i>	33
4.2.3	<i>Árvore de Decisão</i>	34
4.2.4	<i>Gaussian Naive Bayes</i>	36
4.2.5	<i>Máquinas de Vetores Suporte</i>	36
4.3	<i>Seleção dos algoritmos</i>	37
4.4	<i>Métodos de amostragem</i>	37
4.4.1	<i>Métricas de avaliação de desempenho</i>	38
5	Trabalhos Relacionados	43

5.1	<i>Uso de crowdsourcing em problemas de cidades inteligentes</i>	43
5.2	<i>Confiabilidade de crowdsourcing</i>	46
6	Análise de confiabilidade de dados de <i>crowdsourcing</i> com Aprendizado de Máquina	48
6.1	<i>Descrição do modelo de análise da confiabilidade</i>	48
6.2	<i>Conjunto de dados</i>	51
6.3	<i>Coleta de Dados</i>	53
6.4	<i>Análise dos dados</i>	54
6.4.1	Consolidação dos dados	54
6.4.2	Estruturação e normalização dos dados	56
6.4.3	Classificação dos dados	56
6.5	<i>Comparação entre os modelos de análise de confiabilidade</i>	58
6.5.1	Seleção dos método de amostragem	59
6.5.2	Ajuste de Parâmetros	60
6.5.3	Avaliação dos Classificadores	61
7	Resultados e discussões	62
7.1	<i>Análise dos resultados</i>	62
7.1.1	Modelo com Regressão Logística	62
7.1.2	Modelo com <i>K-Nearest Neighbour</i>	64
7.1.3	Modelo com Árvore de Decisão	65
7.1.4	Modelo com <i>Gaussian Naive Bayes</i>	66
7.1.5	Modelo com Máquinas de Vetores Suporte	67
7.2	<i>Comparação dos resultados</i>	69
8	Conclusão	73
8.1	<i>Contribuições</i>	74
8.2	<i>Trabalho publicado</i>	75
8.3	<i>Trabalhos futuros</i>	75
	Referências¹	76

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

APÊNDICES	82
-----------	----

Apêndice A – Exemplo de apêndice	83
--	----

1 Introdução

Cidades Inteligentes (SC — *Smart Cities*) é um conceito utilizado para descrever cidades que promovem a integração de diversos sistemas distribuídos (tais como, câmeras de segurança, sistemas para monitoração de veículos e de trânsito, etc.), para solucionar problemas ocasionados pelo crescimento populacional e, desta forma, criar respostas inovadoras e alinhadas às necessidades da população (SANTIS et al., 2014; BATISTA et al., 2016). Uma característica importante relacionada a SC que requer respostas inovadoras é a *mobilidade urbana*, que poderia ser melhorada com a disponibilização de serviços de transporte coletivo de qualidade (FIGUEIREDO, 2016). Esses serviços influenciam na qualidade de vida da população e têm como objetivo amenizar o impacto ocasionado pelo crescimento populacional (TOWNSEND, 2013).

Um problema recorrente do transporte público é a qualidade dos serviços prestados, especialmente quando se refere ao modal ônibus como meio de transporte coletivo. Em particular, nas grandes metrópoles, as prestadoras de serviços têm dificuldades para estabelecer e cumprir horários de itinerários das linhas de ônibus. A falta de um sistema confiável e previsível leva o usuário a não optar pela utilização desses serviços, o que agrava problemas urbanos sociais e ambientais.

As iniciativas em SC têm como objetivo melhorar a qualidade e gerenciamento dos serviços públicos disponibilizados para a população. Sendo assim, essas iniciativas utilizam dados coletados por meio de diversos sistemas distribuídos, pertencentes as estruturas tecnológicas utilizadas por essas cidades como, por exemplo, os Sistemas de Transporte Inteligentes (ITS — *Intelligent Transport System*), que realizam a coleta de diversas informações por meio de sensores (equipamentos de localização de veículos, semáforos, etc.) (MAGALHÃES, 2008).

No entanto, um dos grandes problemas apresentados por esses sistemas refere-se às limitações decorrentes do uso de dispositivos físicos acoplados aos meios de transporte, tais como, indisponibilidade do sinal GPS ou falha na conexão 3G/4G, que contribuem para que parte dos dados coletados pelos ITS não possa ser utilizada (JEONG; RILETT, 2004; ZHU et al., 2011; PONS; MONTEIRO; CARDOSO, 2015). Para solucionar esses problemas, investiga-se também (STEENBRUGGEN; TRANOS; NIJKAMP, 2015) o uso de novas

tecnologias e abordagens que possam ser aplicadas nas atividades de monitoramento social e urbano como, por exemplo, o *crowdsourcing* (colaboração coletiva).

O *crowdsourcing* é uma metodologia que utiliza dados disponibilizados por um grupo de pessoas para solucionar problemas de um determinado seguimento (BRABHAM, 2013; PEDERSEN et al., 2013). Essa técnica apresenta um imenso potencial para complementar as abordagens tradicionais de levantamento de informações, pois, permitem a obtenção de novos dados ou até mesmo a correção de informações existentes (MISRA et al., 2014; CEROTTI et al., 2016).

1.1 *Motivação e justificativa*

Os dados obtidos com *crowdsourcing* podem ser utilizados para complementar ou até mesmo para realizar a correção de informações existentes sobre o serviço de transporte público. Informações sobre a localização dos ônibus disponibilizadas voluntariamente pelos próprios passageiros podem ser combinadas às informações obtidas pelos dispositivos acoplados aos ônibus de uma rede de transporte público, contribuindo, assim, para a criação de ITS mais eficientes e confiáveis (MISRA et al., 2014; FIGUEIREDO, 2016).

Contudo, os dados de *crowdsourcing* podem ser inerentemente subjetivos e/ou imprecisos, além de potencialmente incorretos (MISRA et al., 2014). É necessário, portanto, que cada nova informação obtida via *crowdsourcing* seja analisada e sua confiabilidade assegurada antes de ser incorporada à base de dados de um ITS.

1.2 *Objetivo e hipótese*

Esta dissertação de mestrado tem por objetivo propor, implementar e validar uma metodologia que combina um conjunto de métricas escolhidas para o contexto do modal ônibus para realizar a análise da confiabilidade dos dados obtidos com *crowdsourcing* e identificar de forma eficiente registros que, por algum motivo, apresentam informações não confiáveis (ausentes ou erradas).

A seguir estão listados os objetivos específicos necessários para alcançar o propósito desta pesquisa:

- I. realizar um levantamento do estado da arte acerca dos temas: *(i)* cidades inteligentes com ênfase na utilização de técnicas de *crowdsourcing*, *(ii)* sistemas inteligentes de transporte e *(iii)* sistemas de informações ao usuário;
- II. realizar um levantamento sobre os principais métodos e abordagens propostas na literatura para a análise de confiabilidades de dados obtidos por *crowdsourcing* e,
- III. desenvolver um modelo que possibilite realizar, de forma automatizada, a análise da qualidade e confiabilidade de um conjunto de dados coletados (com *crowdsourcing*).

A hipótese que será avaliada por esse trabalho é: “Utilizando aprendizado de máquina é possível construir um modelo de análise de confiabilidade, capaz de assegurar a qualidade de dados obtidos voluntariamente com o auxílio de *crowdsourcing* sobre a localização de ônibus de uma rede de transporte público de forma eficiente, com um bom desempenho para ser implementada em uma aplicação em tempo real”.

1.3 Estrutura do documento

Nesse capítulo introdutório foi apresentado, em linhas gerais, o contexto no qual este trabalho de mestrado está inserido, bem como a relevância do problema tratado.

No Capítulo 2 são apresentados os aspectos conceituais relacionados a cidades inteligentes e como tais iniciativas têm colaborado para amenizar os problemas urbanos sociais e ambientais que afetam significativamente a qualidade de vida da população. O Capítulo 3 descreve os principais conceitos relacionados a *crowdsourcing*, e como a sua utilização pode contribuir para a obtenção e análise de diversos tipos de informações que normalmente não poderiam ser capturadas por meio da utilização de abordagens e técnicas tradicionais. Uma breve introdução sobre as principais técnicas e conceitos relacionados a aprendizado de máquina é apresentada no Capítulo 4. Esses capítulos descrevem ainda, como tais abordagens podem ser utilizadas para desenvolver um modelo de análise de confiabilidade de dados de *crowdsourcing* especializado para o sistema de transporte público do modal ônibus, fornecendo também algumas diretrizes importantes para o entendimento dos assuntos tratados nos capítulos posteriores.

No Capítulo 5, são apresentados os trabalhos encontrados na literatura que abordam a temática do uso de *crowdsourcing* em problemas de cidades inteligentes. Já o Capítulo 6 discute o modelo proposto em um nível maior de profundidade, descrevendo

os procedimentos de coleta, análise e tratamento do conjunto de dados, bem como as abordagens e métodos empregados no desenvolvimento desse trabalho de mestrado. O Capítulo 7 apresenta detalhadamente os resultados obtidos por meio da execução dos experimentos conduzidos por essa pesquisa.

Por fim, no Capítulo 8 são apresentadas as considerações finais sobre os resultados alcançados, e as perspectivas para trabalhos futuros.

2 Cidades inteligentes, mobilidade urbana e suas tecnologias

Este capítulo apresenta os principais conceitos inerentes às cidades inteligentes, bem como descreve alguns domínios e visões acerca do tema. A Seção 2.1 apresenta os conceitos de cidades inteligentes, que visa melhorar a qualidade dos serviços disponibilizados para a população, por meio do uso inteligente de tecnologias. Na Seção 2.2 discorremos sobre os conceitos de transportes e como eles influenciam na qualidade de vida dos habitantes das grandes cidades. A Seção 2.3 apresenta uma introdução aos conceitos de sistemas inteligentes de transportes e como eles podem contribuir para melhorar a qualidade dos serviços de transporte públicos. Por fim, a Seção 2.4 apresenta os conceitos das aplicações de sistemas de informação ao usuário.

2.1 Cidades inteligentes

O conceito de Cidades Inteligentes (*SC — Smart Cities*) engloba iniciativas com o objetivo de atenuar e gerir eficientemente problemas urbanos que afetam consideravelmente a qualidade de vida da população, decorrentes do constante crescimento populacional, tais como: deficiências na gestão de resíduos; escassez, desperdícios e má gestão dos recursos naturais; restrições nos sistemas de saúde, educação e segurança pública; limitações nos sistemas de mobilidade urbana e de transportes; obsolescência e encurtamento do ciclo de vida das infraestruturas públicas (SHAPIRO, 2006; TOPPETA, 2010; BATAGAN, 2011; CAPROTTI, 2016; BATISTA et al., 2016). Esse termo tem sido utilizado para descrever cidades que buscam continuamente o desenvolvimento urbano, mediante políticas que estimulam a participação da população por meio de inteligências distintas, como a *(i)* humana, *(ii)* coletiva e *(iii)* artificial (por meio das TIC — Tecnologias da Informação e Comunicação) (CURY; MARQUES, 2016).

No contexto de *mobilidade urbana*, um dos principais objetivos das SC é resolver os problemas relacionados à disponibilização de serviços de transporte de qualidade para a população (FIGUEIREDO, 2016).

2.2 Conceitos sobre transporte urbano

O transporte urbano pode ser definido como o processo de deslocamento de pessoas (transporte de passageiros) e/ou produtos (transporte de cargas). Os meios de transporte de passageiros afetam diretamente a qualidade de vida da população, porque definem as alternativas que os habitantes têm a sua disposição para se locomover até os locais de trabalho, áreas de lazer, locais de atendimento à saúde, centros culturais, entre outros (FERRAZ; TORRES, 2004). O Quadro 1 apresenta as categorias, características e modos comumente utilizados nos deslocamentos diários da população.

Quadro 1 – Classificação dos modos de transporte urbano de passageiros.

MODOS	DESCRIÇÃO	TRANSPORTES
Privado ou individual	Os veículos são conduzidos por um dos usuários, que pode escolher livremente o caminho e o horário de partida. Há, portanto, total flexibilidade de uso no espaço e no tempo. A capacidade do veículo é pequena e a posse pode ser momentânea.	A pé, bicicleta, motocicleta e automóvel próprio.
Público, coletivo ou de massa	Os veículos pertencem, em geral, a uma empresa e operam em rotas predefinidas e horários fixos. Não há flexibilidade de uso no espaço e no tempo. A capacidade do veículo é grande.	Ônibus, bonde, metrô e trem.
Semi-público	O veículo pertence a uma empresa ou indivíduo e pode ser utilizado por determinado grupo de indivíduos ou por qualquer pessoa, tendo rota e horários adaptáveis aos desejos dos usuários em vários graus.	Táxi, Uber, moto-táxi, carona programada, lotação, veículo fretado ou alugado.

Fonte: Ferraz e Torres (2004), adaptado pelo autor.

A utilização dos serviços de transporte público contribui para a redução da emissão de gases poluentes (devido ao menor número de veículos em vias públicas) e do índice de tráfego nas grandes cidades (FERRIS; WATKINS; BORNING, 2010). Tais serviços podem ser divididos em duas categorias: *(i)* sobre pneus (ônibus, micro-ônibus, entre outros) e *(ii)* sobre trilhos (metrô, trem e etc.) (FERRAZ; TORRES, 2004). Entretanto, existem alguns problemas relacionados a transporte público que precisam ser tratados, como, por exemplo, a falta de informações sobre os serviços disponibilizados para os usuários (TOPPETA, 2010; BATAGAN, 2011).

Conforme destacado na Figura 1, a cidade de São Paulo apresenta um dos maiores e mais complexos sistemas de transporte de ônibus do mundo. A cidade possui mais de

14.500 ônibus no âmbito municipal, os quais circulam em mais de 1.350 linhas (PONS; MONTEIRO; CARDOSO, 2015). Esses serviços são disponibilizados por 37 empresas que realizam cerca de 195 mil viagens diárias, o que representa um total de mais de 2,4 bilhões de pessoas transportadas por ano.

Figura 1 – Quantidade de ônibus e linhas em algumas cidades do mundo.

Fonte: Pons, Monteiro e Cardoso (2015).

Segundo a Associação Nacional de Transportes Públicos (ANTP), estima-se que em 2050 aproximadamente 70% da população do mundo viverá em áreas urbanas (ANTP, 2016). Estes dados geram um grande desafio relacionado à infraestrutura das cidades e requer um planejamento estratégico do transporte público urbano (HANDWERK, 2008).

O uso de tecnologias inovadoras no setor de transporte público contribuem para aumentar a qualidade dos serviços de transportes oferecidos e, deste modo, atrair novos usuários (SUSSMAN, 2008; MISRA et al., 2014; CEROTTI et al., 2016). A integração eficiente das tecnologias existentes, tem permitido a criação e identificação de inúmeras oportunidades para solucionar os principais problemas que afetam as grandes cidades. Os sistemas de transportes inteligentes, podem ser descritos como um exemplo de tecnologia que possibilita o monitoramento de tráfego e sua manutenção preventiva, bem como, a possibilidade de localização e georreferenciamento de transportes alternativos e informações sobre itinerários (WOLFRAM, 2012).

2.3 *Sistemas Inteligentes de Transporte*

Os Sistemas de Transportes Inteligentes (ITS — *Intelligent Transportation Systems*) são programas responsáveis por coletar informações sobre os veículos de uma rede de transporte público e utilizar esses dados em aplicações inteligentes (FIGUEIREDO, 2016). Os ITS são utilizados para: (i) identificar a localização dos veículos através de sensores ou GPS, (ii) transmitir, receber e processar grandes quantidades de dados e (iii) utilizar essas informações para melhorar as condições do tráfego (SUSSMAN, 2008).

Por meio de sensores (equipamentos de localização de veículos, semáforos, etc.), esses sistemas realizam a coleta de diversos tipos de dados. Essas informações podem ser utilizadas para gerar conhecimento que auxilie os gestores públicos no planejamento eficiente dos serviços prestados, como, por exemplo, para melhorar o gerenciamento de tráfego e fornecer sugestões de rotas aos motoristas. Esses sistemas também permitem disponibilizar informações aos usuários sobre condições meteorológicas, condições de estradas, possíveis perigos e itinerário de veículos de uma rede transporte público (MAGALHÃES, 2008; GOMEZ; DIAZ; BOUSSETTA, 2009).

Um grupo de ITS relevantes para esse trabalho são os Sistemas Avançados de Transporte Público (APTS — *Advanced Public Transportations Systems*), que podem ser descritos como um conjunto de aplicações que aumentam a qualidade e a segurança dos sistemas de transporte públicos, de modo a disponibilizar um serviço mais eficiente para seus respectivos usuários (FIGUEIREDO, 2005). Os APTS utilizam um sistema de Localização Automática de Veículos (AVL — *Automatic Vehicle Location*), que possibilita o rastreamento dos veículos em tempo real (CHOWDHURY; SADEK, 2003).

Os sistemas AVL podem ser utilizados para diversas finalidades, tais como, no gerenciamento de transportes públicos, monitoração de veículos oficiais (viaturas policiais e hospitalares), ou ainda para propósitos militares (CHOWDHURY; SADEK, 2003). Esses sistemas são compostos por duas tecnologias importantes responsáveis por realizar a (i) captura da localização e a (ii) transmissão de dados, conforme é ilustrado na Figura 2 (PONS; MONTEIRO; CARDOSO, 2015). A primeira utiliza um sistema de posicionamento global (GPS — *global positioning system*) para capturar a localização atual do veículo. Por sua vez, a segunda realiza a transmissão dos dados coletados a uma central de controle e monitoramento (por intermédio de uma conexão 3G/4G), onde,

posteriormente, essas informações são analisadas (CUNHA, 2008; PONS; MONTEIRO; CARDOSO, 2015).

Figura 2 – Arquitetura AVL (*Automatic Vehicle Location*).

Fonte: Pons, Monteiro e Cardoso (2015)

Contudo, assim como diversas tecnologias, os sistemas AVL apresentam limitações. Por exemplo, os sinais de satélite utilizados pelo sistema GPS não alcançam o subsolo e podem ser interrompidos por inúmeros fatores ambientais, como, localidades onde existe uma grande quantidade de edifícios ou zonas muito arborizadas (JEONG; RILETT, 2004; PONS; MONTEIRO; CARDOSO, 2015).

A utilização de conexões 3G/4G também é uma fonte de problemas, uma vez que a cobertura das operadoras de telefonia varia de acordo com a região. Deste modo, a própria arquitetura AVL faz com que uma parte dos dados coletados não seja explorável (PONS; MONTEIRO; CARDOSO, 2015).

2.4 Sistemas de informação ao usuário

Incorporado aos APTS encontram-se os Sistemas de Informação ao Usuário (SIU). Por meio desses sistemas, é possível fornecer aos passageiros informações relevantes, tais como: (i) horários de chegada, (ii) tempo de espera, (iii) localização do próximo veículo e (iv) sugestões de rotas (PILON, 2009). Com a disponibilização desses dados, há a possibilidade de que os usuários realizem um planejamento melhor de suas viagens, economizando tempo e causando menor estresse, ao mesmo tempo que se incentiva o usuário

a utilizar o transporte público ao invés do privado (PONS; MONTEIRO; CARDOSO, 2015).

Para disponibilizar essas informações, os SIU fazem uso de tecnologias como dispositivos móveis, monitores, painéis eletrônicos e computadores (SILVA, 2000). O Quadro 2 destaca as principais funções desses sistemas. Na cidade de São Paulo, desde 2008, os usuários contam com informações instantâneas fornecidas pelo SIU chamado de Sistema Integrado de Monitoramento (SIM), também conhecido como *Sistema Olho Vivo* (SPTRANS, 2009).

Quadro 2 – Funções dos Sistemas de Informação ao Usuário.

Promocional	<ul style="list-style-type: none"> • Mobilidade: propor motivos para viagens e possíveis destinos; • Presença: informar as pessoas sobre o transporte público como parte do pacote de facilidades ofertadas; • Imagem: melhorar a imagem do transporte público.
Ensinamento	<ul style="list-style-type: none"> • Entendimento: informar como utilizar o transporte público; • Adequabilidade: divulgar as regras envolvidas no uso dos sistemas.
Operacional	<ul style="list-style-type: none"> • Planejamento de Viagens: informar sobre restrições e oportunidades associadas com o uso do sistema para diferentes tipos de viagens; • Acesso: capacitar pessoas para acesso à rede de transporte público; • Viagem: capacitar a realização de uma viagem; • Modificação: informar sobre mudanças na programação.
Moderação	<ul style="list-style-type: none"> • Comportamento: aliviar a ansiedade do viajante; • Controle: aumentar o controle do usuário sobre a escolha entre as opções disponíveis.

Fonte: SILVA (2000), adaptado pelo autor.

Entretanto, as limitações das tecnologias utilizadas pelos SIU, tais como indisponibilidade do sinal de GPS ou falha na conexão 3G/4G, contribuem para que uma parte importante dos dados não seja utilizada (ZHU et al., 2011; PONS; MONTEIRO; CARDOSO, 2015). Para solucionar esses problemas, investiga-se o uso de outras abordagens de obtenção de dados (STEENBRUGGEN; TRANOS; NIJKAMP, 2015). Neste sentido, as técnicas de *crowdsourcing* apresentam um imenso potencial para complementar as abordagens tradicionais e são perfeitamente aplicáveis no cenário de transporte público (MISRA et al., 2014; CEROTTI et al., 2016).

3 Conceitos e tecnologias relacionados a crowdsourcing

Este capítulo apresenta os principais conceitos relacionados às técnicas de *crowdsourcing*. A Seção 3.1 introduz as principais características de um modelo de *crowdsourcing*, mostrando como esse método pode ser utilizado para realizar a captura de diversas informações. A Seção 3.2 descreve como a utilização dessas abordagens pode contribuir de forma significativa para o desenvolvimento de SIUs mais eficientes. Por fim, a Seção 3.3 descreve o aplicativo Coletivo, ferramenta utilizada para coletar parte do conjunto de dados utilizado por essa pesquisa.

3.1 Crowdsourcing

O termo *crowdsourcing* (*colaboração coletiva* ou *contribuição colaborativa*) se refere ao uso de uma rede distribuída de voluntários dispostos a resolver problemas, desenvolver novas tecnologias, contribuir com dados, etc. (HOWE, 2006). Essa metodologia utiliza dados disponibilizados por meio de diversas plataformas (como, por exemplo, fóruns, sites, aplicativos para dispositivos móveis, etc.), para gerar informações relevantes que auxiliem no processo de tomada de decisão de um determinado seguimento (BRABHAM, 2013; PEDERSEN et al., 2013; MISRA et al., 2014; SCHWARTZ et al., 2015).

Esse termo tem sido discutido como uma abordagem inovadora para compartilhar conhecimento e engajar o público de diversos seguimentos. As iniciativas baseadas nessa metodologia estão se tornando essenciais para as SC, uma vez que possibilitam a captura de informações que normalmente não poderiam ser coletadas por meio da utilização de abordagens e técnicas tradicionais (CULLINA; CONBOY; MORGAN, 2015).

Um dos principais desafios das iniciativas de SC é atenuar os problemas referentes à mobilidade urbana. Diversos sistemas distribuídos e abordagens de levantamento de informações têm sido propostos e utilizados para solucionar tais problemas. Sendo assim, *crowdsourcing* apresenta uma solução inovadora com um imenso potencial para complementar tais iniciativas, pois permite a obtenção de novos dados (MISRA et al., 2014; CEROTTI et al., 2016).

As informações coletadas com *crowdsourcing*, em tese, podem ser utilizadas para manter as bases de dados atualizadas e gerar conhecimentos que auxiliem no planeja-

mento e gerenciamento dos serviços oferecidos para a população, ao mesmo tempo, que disponibilizam uma melhor visibilidade dos problemas para os gestores/prestadores de serviços públicos (MISRA et al., 2014; SCHWARTZ et al., 2015; CEROTTI et al., 2016). Esses dados, podem ser utilizados, por exemplo, para a criação de SIUs mais eficientes e confiáveis, proporcionando deste modo, uma melhor interação entre usuários e gestores de serviços de transporte públicos (PEDERSEN et al., 2013; SCHWARTZ et al., 2015).

Contudo, assim como as abordagens tradicionais, tipicamente propensas a erros devido às limitações das tecnologias utilizadas, os dados disponibilizados pelos usuários podem ser inerentemente subjetivos e/ou imprecisos, e potencialmente incorretos. Portanto, a extração e análise desses dados é uma atividade que requer o desenvolvimento de técnicas específicas para esse cenário (MISRA et al., 2014).

Deste modo, considerando a necessidade de assegurar a confiabilidade das informações obtidas com *crowdsourcing*, esse trabalho propõe o uso de técnicas de aprendizado de máquina para criar um modelo que realize de forma automática a análise da confiabilidade dos dados e, desta forma, mitigar as incertezas introduzidas pelo uso de *crowdsourcing*.

3.2 *Crowdsourcing em sistemas de informação ao usuário*

A utilização de dispositivos móveis como ferramenta para a aquisição e disponibilização de informações em tempo real tem sido amplamente explorada por pesquisadores. O aumento da capacidade computacional, juntamente com a integração de microfone, câmera, GPS, acelerômetro, conexão de internet e etc desses dispositivos, em conjunto com as metodologias de *crowdsourcing*, representam um novo paradigma no suporte à coleta de dados (BRABHAM, 2013; MISRA et al., 2014).

As possibilidades proporcionadas pelo uso de dispositivos móveis, aliada à sua popularidade entre os usuários, possibilitam que essa tecnologia torne-se uma poderosa ferramenta de coleta de dados (PEDERSEN et al., 2013; SCHWARTZ et al., 2015; CULLINA; CONBOY; MORGAN, 2015). As informações sobre a localização de um ônibus disponibilizadas voluntariamente pelos próprios usuários a partir de seus próprios dispositivos, podem ser combinadas, por exemplo, às informações obtidas pelos sistemas

AVL acoplados aos veículos de uma rede de transporte público, contribuindo, assim, para a criação de SIUs mais eficientes e confiáveis (MISRA et al., 2014; FIGUEIREDO, 2016).

Além disso, o uso de tais de recursos facilita a comunicação entre os usuários e os gestores/prestadores de serviços. A sua utilização possibilita que os usuários contribuam com informações relevantes e em tempo real sobre os problemas vivenciados em seu cotidiano. Isso proporciona uma melhor visibilidade dos problemas para os gestores/prestadores de serviços públicos (MISRA et al., 2014; SCHWARTZ et al., 2015).

Existem algumas iniciativas voltadas para o desenvolvimento de SIUs, especificamente para dispositivos móveis. Tais sistemas fornecem inúmeras informações para os seus usuários. A seguir serão apresentados, brevemente, alguns dos exemplos de sistemas relacionados com esta pesquisa.

3.2.1 Moovit

O Moovit¹ é um aplicativo que permite a troca de informações, em tempo real, sobre o transporte público nas principais cidades das Américas do Norte e do Sul, Europa e Ásia. Esse aplicativo disponibiliza informações que possibilitam a seus usuários planejar suas viagens utilizando quaisquer tipos de transporte público (PAGANI et al., 2016). O Moovit combina informações do transporte público com informações de tráfego em tempo real coletadas dos próprios usuários. Com atualizações constantes sobre as condições do trânsito, o aplicativo mostra a melhor rota, estima o tempo de chegada, o momento de desembarque do ônibus e troca de modais, possibilitando a classificação das linhas com base em parâmetros como pontualidade e segurança no trajeto (TORREGROSA; IGARTUA; PUGLISI, 2016).

3.2.2 CityMapper

CityMapper² é uma aplicação para transporte urbano que combina informações de diferentes mapas em tempo real (Apple Maps, Google Apps, Open Street Map, sites dos sistemas de transportes locais, previsão do tempo e de trânsito) com os dados disponibilizados pelos seus usuários por meio de *crowdsourcing*. O aplicativo utiliza as informações

¹ <https://moovitapp.com>

² <https://citymapper.com>

coletadas para encontrar o trajeto mais rápido, barato ou fácil de transporte público ou privado, dependendo dos requisitos de cada usuário. O CityMapppper está disponível apenas para grandes cidades, como Londres, Nova Iorque, Paris, Berlim, São Paulo, México e Singapura (SAMSEL et al., 2014; PAGANI et al., 2016).

3.2.3 OneBusAway

O OneBusAway³ é uma ferramenta que disponibiliza informações em tempo real sobre o ônibus no estado de Washington, nos Estados Unidos. Esse sistema informa aos seus usuários os horários de chegada, bem como, os próximos pontos de paradas e se há atraso ou não em uma determinada linha (FERRIS; WATKINS; BORNING, 2009). As informações em tempo real sobre a movimentação dos veículos utilizadas pelo OneBusAway são obtidas de uma entidade pública externa, sendo somente tratadas e processadas para serem utilizadas no sistema. Além disso, o sistema utiliza *crowdsourcing* para coletar informações contextuais ligadas à localização do usuário para apresentar alertas sobre problemas em seu trajeto (FERRIS; WATKINS; BORNING, 2011; BARBEAU; BORNING; WATKINS, 2014).

3.3 Descrição do aplicativo Coletivo

Uma iniciativa de SIU importante para esse trabalho é o aplicativo Coletivo⁴, um sistema para dispositivos móveis desenvolvido pela *startup* Scipopulis⁵, uma empresa de inovação para cidades inteligentes e dedicada à mobilidade urbana, financiada pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Esse aplicativo fornece aos seus usuários informações em tempo real sobre os ônibus da rede de transportes público da cidade de São Paulo.

Conforme ilustra a Figura 3, o aplicativo apresenta uma funcionalidade que permite aos usuários informarem quando um determinado ônibus passou em um dos pontos de parada. Quando o passageiro informa a localização de um ônibus o sistema realiza a

³ <https://onebusaway.org>

⁴ <http://appcoletivo.com>

⁵ Scipopulis: <http://scipopulis.com/>

captura de informações como horário, localização, ponto de parada e número da linha do ônibus.

Figura 3 – Configurações do aplicativo Coletivo.

Fonte: Elaborado pelo autor.

A coleta de dados realizada por um SIU como o Coletivo permite a obtenção de novas informações que podem ser utilizadas para enriquecer a base de dados utilizada por esses aplicativos, contribuindo, assim, para a criação de sistemas mais eficientes e confiáveis. Entretanto, esses dados também podem apresentar informações incorretas. Portanto, é necessário que cada nova informação coletada seja analisada e sua confiabilidade assegurada antes de ser incorporada à base de dados.

Deste modo, esse trabalho propõe o uso de técnicas de aprendizado de máquina para criação de modelos que realize de forma automática a análise da confiabilidade dessas informações. Parte dos dados utilizado para desenvolver esses modelos, foi coletado utilizando o aplicativo Coletivo.

4 Conceitos e técnicas relacionadas a Aprendizado de Máquina

Conforme descrito no capítulo anterior, os dados obtidos com *crowdsourcing* podem apresentar informações não confiáveis. Para resolver esse problema, este trabalho propõe um modelo que utiliza técnicas de aprendizado de máquina para identificar de forma eficiente registros que, por algum motivo, apresentam informações que possam comprometer a qualidade dos dados coletados. Sendo assim, neste capítulo é feita uma introdução aos conceitos de aprendizado de máquina, tema ao qual está relacionado este projeto de mestrado. Na Seção 4.1 são apresentadas algumas considerações iniciais relacionadas a aprendizado de máquina. A Seção 4.2 descreve o método de aprendizado supervisionado, bem como, os algoritmos classificadores considerados no desenvolvimento dos modelos de análise de dados de *crowdsourcing*. Já a Seção 4.3 apresenta as principais técnicas utilizadas para realizar a avaliação de desempenho dos modelos desenvolvidos.

4.1 Aprendizado de Máquina

Aprendizado de Máquina (ML — *Machine Learning*) é um método que utiliza conceitos de inteligência artificial e/ou técnicas estatísticas para realizar o reconhecimento de padrões (MITCHELL, 1997). Esse método é utilizado no desenvolvimento de sistemas inteligentes capazes de realizar, automaticamente, a identificação de padrões em um conjunto de dados e utilizar os padrões descobertos para prever acontecimentos futuros, ou até mesmo para realizar outros tipos de tomada de decisão (MURPHY, 2012; ALPAYDIN, 2014).

As técnicas de ML são aplicadas em várias áreas de pesquisa, como, por exemplo, sistemas especialistas, compreensão da linguagem natural, robôs inteligentes e em aplicações específicas, como motores de busca, diagnósticos médicos, reconhecimento facial, detecção de fraudes, entre outros (GHOTRA; MCINTOSH; HASSAN, 2015). Os algoritmos de ML podem ser classificados em 3 categorias (MARSLAND, 2015):

- **Aprendizado supervisionado:** esse método utiliza um conjunto de dados (*amostra*) que contém registros de exemplos com suas respectivas classes (*rótulos*). Baseado nesses exemplos, o algoritmo classificador generaliza as respostas recebidas e aplica

essas generalizações a registros novos e de classes desconhecidas, com a finalidade de determinar corretamente suas respectivas classes.

- **Aprendizado não supervisionado:** esse método não utiliza um conjunto de dados contendo registros de exemplo com classes conhecidas. Ao invés disso, o algoritmo tenta identificar similaridades entre os registros para atribuir a mesma classe aos registros parecidos.
- **Aprendizado por reforço:** esse método pode ser descrito como um tipo de aprendizado situado no meio-termo entre os tipos anteriores. Neste tipo de aprendizado, o algoritmo é informado apenas quando um registro é erroneamente classificado. Desta forma, o algoritmo tem que aprender por si só, explorando algumas alternativas até encontrar a correta.

4.2 *Aprendizado supervisionado*

Neste trabalho de dissertação será utilizado o método de aprendizado supervisionado. Esse método utiliza algoritmos de classificação que são capazes de realizar aprendizagem interativa, mediante a análise de um conjunto de dados chamados de *amostra* (BISHOP, 2006; GHOTRA; MCINTOSH; HASSAN, 2015). O classificador, ou método de classificação, é responsável por fazer as atribuições entre os elementos de um conjunto de dados (*dataset*¹) e os elementos das suas respectivas classes (MITCHELL, 1997). O processo de classificação realizado pelo método de aprendizado supervisionado pode ser descrito pelo par ordenado (X, y) , onde X é um vetor que representa os atributos de um registro, $X = (x_1, x_2, x_3, \dots, x_n)$, e y é o rótulo da classe a qual este registro pertence (MITCHELL, 1997; MARSLAND, 2015). A Figura 4 apresenta uma esquematização deste processo.

A Tabela 1 descreve um exemplo de *dataset* que contém parte dos registros coletados e considerados para realizar a análise da confiabilidade das informações sobre a localização dos ônibus da rede de transporte público, obtidas mediante a utilização de técnicas de *crowdsourcing*. Cada linha representa um registro do conjunto de dados e cada coluna um atributo desse registro. O atributo *Classe* representa o rótulo da classe atribuída para cada registro, ou seja, 1 para “*Confiável*” e 0 para “*Não Confiável*”.

¹ O termo *dataset* é usado como sinônimo de conjunto de dados (amostra) neste trabalho, por se tratar de um termo amplamente difundido em ML.

Figura 4 – Representação do processo de classificação.

Fonte: Baseado em Mitchell (1997) e Marsland (2015).

Tabela 1 – Exemplo de um amostra dos dados de *crowdsourcing*.

Exemplos / Registros		Atributos (X)			Classe (y)
Nº linha	Nome da linha	Distância usuário (M)	Distância anterior (M)	Velocidade anterior (Km/h)	Confiável / Não Confiável ⁽¹⁾
172U-10	Cem. Pq. Dos Pinheiros	105,29	38,99	10,03	1
5154-10	Term. Sto. Amaro	1634,77	1301,55	23,66	0
648P-10	Term. Pinheiros	97,98	306,25	28,27	1
476A-10	Term. Sto. Amaro	99,99	525,93	32,09	1
6013-10	Term. Sto. Amaro	98,5	453,08	36,25	1
669A-10	Term. Princ. Isabel	1302,95	241,22	37,76	0
775V-10	Itaim Bibi	97,65	584,46	43,83	1
509J-10	Pq. Ibirapuera	4941,53	844,46	434,29	0
857P-10	Term. Campo Limpo	1306,67	14496,57	660,6	0

1) 0 = Não Confiável 1 = Confiável

O desenvolvimento de um classificador passa por duas fases: (i) treinamento e (ii) teste. A primeira fase é utilizada para criar uma função que tenta descrever os dados em relação às classes do problema. Já a segunda consiste em utilizar a função gerada na fase de treinamento para verificar em qual das classes um novo registro se enquadra (MITCHELL, 1997). Para assegurar uma boa capacidade de generalização, ou seja, a habilidade de classificar corretamente novos registros, é recomendável que os dados aplicados na fase de teste não sejam os mesmos utilizados durante o treinamento (MITCHELL, 1997; BISHOP, 2006; GHOTRA; MCINTOSH; HASSAN, 2015; MARS LAND, 2015).

Quando um classificador apresenta uma baixa capacidade de generalização, dizemos que ele sofreu um superajustamento aos dados de treinamento. Esse superajuste é

denominado de *overfitting*. O caso contrário ao *overfitting* é denominado de *underfitting*, e ocorre quando o classificador gerado pelo algoritmo não possui uma boa capacidade de classificação mesmo com o conjunto de treinamento (MITCHELL, 1997; MARSLAND, 2015).

Figura 5 – Processo de classificação e sua utilização na predição de novos exemplos.

Fonte: Baseado em Mitchell (1997) e Marsland (2015).

A Figura 5 descreve o processo de classificação realizado pelos algoritmos de aprendizado supervisionado e posteriormente a sua utilização. Inicialmente o *dataset* no qual os rótulos das classes dos registros são conhecidos é utilizado para construir e treinar o modelo. Em seguida, esse modelo é aplicado para prever os rótulos dos dados cujas classes não são conhecidas.

Diversos algoritmos de ML podem ser utilizados para desenvolver um modelo de classificação (GHOTRA; MCINTOSH; HASSAN, 2015). Para este estudo, foram considerados os algoritmos frequentemente utilizados: Regressão Logística; *K-Nearest Neighbour*; Árvore de Decisão; *Gaussian Naive Bayes* e Máquinas de Vetores Suporte.

4.2.1 Regressão Logística

A regressão logística é uma técnica estatística utilizada para entender a relação entre uma variável resposta (Y) e uma ou mais variáveis independentes ($X = (x_1, x_2, \dots, x_p)$) e permite construir um modelo para explicar essa associação. Na regressão logística, a

variável resposta (Y) é dicotômica, isto é, atribui-se o valor 1 para o acontecimento de “sucesso” e 0 para “fracasso” (HOSMER; LEMESHOW; STURDIVANT, 2013). O modelo de regressão logística pode ser descrito matematicamente pela Equação 1.

$$P(Y = 1) = \frac{1}{1+e^{-g(x)}} \quad (1)$$

onde,

$$g(x) = B_0 + B_1X_1 + \dots + B_pX_p$$

Os coeficientes B representam os parâmetros que serão utilizados para determinar o valor estimado para $g(x)$, com $0 \leq g(x) \leq 1$. Assim, sabendo os parâmetros B e conhecendo os valores das variáveis independentes para um determinado registro, podemos aplicar a fórmula acima para calcular a probabilidade de que este registro apresenta informações não confiáveis (HOSMER; LEMESHOW; STURDIVANT, 2013). Logo, a regra de classificação pode ser descrita da seguinte forma:

- Quando $g(x) \rightarrow +\infty$, então $P(Y = 1) \rightarrow 1$,
- Quando $g(x) \rightarrow -\infty$, então $P(Y = 1) \rightarrow 0$,

Com a utilização dos métodos de regressão logística, é possível prever o valor que a variável resposta terá diante dos valores das variáveis independentes, fazendo com que o método possa, diante de algumas circunstâncias, atuar como um classificador (HOSMER; LEMESHOW; STURDIVANT, 2013).

4.2.2 *K-Nearest Neighbour*

O algoritmo *K-Nearest Neighbour* (KNN) é um algoritmo de aprendizado de máquina supervisionado que realiza aprendizagem por analogia, ou seja, esse algoritmo relaciona cada um dos registros do conjunto de dados a um ponto em um espaço n -dimensional, sendo n o número de atributos de entrada que descrevem o conjunto de dados (MITCHELL, 1997). Para classificar um novo registro, a similaridade com outros registros já conhecidos é calculada por meios da distância apresentada por tais registros

(WITTEN et al., 2016). Essa distância, pode ser calculada mediante a utilização da medida de distância euclidiana (MITCHELL, 1997) descrita pela Equação 2.

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2} \quad (2)$$

Onde, x descreve o vetor de características $x = a_1(x), a_2(x), \dots, a_n(x)$ apresentado por um registro, e $a_r(x)$ é utilizado para denotar o valor do r -ésimo atributo do registro x . Deste modo, o valor da distância entre dois registros x_i e x_j pode ser definido por $d = (x_i, x_j)$.

4.2.3 Árvore de Decisão

Árvore de decisão são algoritmos que apresentam uma estrutura no formato de árvore onde cada nó interno representa um teste de uma determinada variável, que define o caminho percorrido da raiz até uma folha da árvore (nó terminal). As folhas da árvore representam as decisões a serem tomadas, ou seja, os resultados da classificação. De modo geral, a predição da variável-alvo é feita através de regras de decisão simples, inferidas pelos dados de treinamento. Esse algoritmo utiliza a estratégia de “*divisão e conquista*” onde, a cada passo, um determinado atributo (que melhor descreve a classe à qual este registro pertence) é selecionado e os dados são particionados de acordo com os valores do atributo (cada valor define um ramo, ou seja, o caminho a ser percorrido na árvore). O processo é repetido recursivamente para os ramos inferiores utilizando somente o subconjunto de dados de cada ramo e as variáveis que sobraram.

A Figura 6 apresenta um exemplo simples de uma Árvore de Decisão, adaptado para o contexto de análise de confiabilidade de dados sobre a localização dos ônibus da rede de transporte público disponibilizados voluntariamente pelos usuários. Neste exemplo, cada retângulo é um teste em um dos atributos para um dado conjunto de dados de *crowdsourcing*. Cada elipse representa uma classe, ou seja, o rótulo atribuído para cada registro do conjunto de dados.

Figura 6 – Árvore de Decisão simples para análise de confiabilidade de dados sobre a localização dos ônibus.

Fonte: Elaborado pelo autor.

Vários algoritmos foram implementados para assegurar a construção de Árvore de Decisão e seu uso para a tarefa de classificação. Os algoritmos ID3 e C4.5 (QUINLAN, 1993) e CART (BREIMAN et al., 1984), são os mais populares e frequentemente utilizados na construção de Árvore de Decisão (DUDA; HART; STORK, 2000).

O ID3 (*Iterative Dichotomiser 3*) é um algoritmo utilizado para classificar dados nominais (não ordenados). Esse algoritmo realiza o cálculo do ganho de informação durante a divisão dos atributos ao longo da árvore. Já C4.5 pode ser descrito como uma versão superior do ID3. Esse algoritmo utiliza um conjunto de heurísticas específicas para podar as árvores geradas não contemplado pelo ID3. O C4.5 permite ainda a utilização de atributos contínuos e categorizados na construção de um modelo de Árvore de Decisão (QUINLAN, 1993; DUDA; HART; STORK, 2000).

O CART (*Classification on Regression Trees*) é um algoritmo que apresenta uma grande capacidade de identificar padrões, mesmo quando eles não são evidentes, bem como, a produção de resultados sob a forma de Árvore de Decisão de grande simplicidade e legibilidade. O resultado deste algoritmo é sempre uma árvore binária que pode ser percorrida da sua raiz até as folhas respondendo apenas a questões simples do tipo “sim” ou “não” (BREIMAN et al., 1984; DUDA; HART; STORK, 2000).

Devido a sua fácil utilização e implementação, o algoritmo de Árvore de Decisão CART foi escolhido para a construção de um dos métodos de análise de dados de *crowd-sourcing* proposto nesta dissertação.

4.2.4 *Gaussian Naive Bayes*

O *Gaussian Naive Bayes* é um classificador probabilístico baseado no *Teorema de Bayes*. Ele é utilizado para calcular a probabilidade da ocorrência de um evento, baseando-se em probabilidades obtidas da análise dos eventos passados (MITCHELL, 1997; WITTEN et al., 2016). Normalmente o algoritmo apresenta uma boa eficiência computacional e robustez. Entretanto, esse método depende muito da distribuição das informações extraídas pelos descritores, pois o classificador elabora sua função de decisão a partir da suposição de que os dados possuem distribuição normal. Então, quanto mais próxima a distribuição for da normal, melhor o desempenho do classificador (WITTEN et al., 2016).

4.2.5 Máquinas de Vetores Suporte

Os algoritmos de máquinas de vetores suporte (SVM — *Support Vector Machine*) podem ser definidos como um conjunto de métodos que utilizam técnicas de classificação para realizar o reconhecimento de padrões em um determinado conjunto de dados (VAPNIK, 1998; HSU; CHANG; LIN, 2010). As SVMs baseiam-se na teoria de aprendizado estatístico (TAE), para realizar a construção de hiperplanos e, desta forma, separar um conjunto de dados em classes distintas (SCHOLKOPF; SMOLA, 2001).

De forma simplificada, o objetivo de uma SVM é encontrar um hiperplano ótimo, capaz de separar o maior número de pontos de uma mesma classe do mesmo lado, maximizando a distância entre os pontos e o hiperplano de separação. Para isso esse algoritmo faz uso de uma função de *kernel*, que possui um conjunto de parâmetros que podem ser manipulados para obtenção de um melhor desempenho do classificador (SCHOLKOPF; SMOLA, 2001; HSU; CHANG; LIN, 2010).

4.3 Seleção dos algoritmos

A escolha do melhor algoritmo para o processo de criação de um modelo de classificação é considerada uma atividade complexa, pois, além dos algoritmos utilizados para criar um classificador, existem diversos fatores (a parametrização que será utilizada, a complexidade computacional do algoritmo, as variáveis descritivas que serão utilizadas, o método de amostragem que será aplicado no conjunto de dados, entre outros) que podem influenciar significativamente no desempenho de um modelo (MITCHELL, 1997; JAMES et al., 2013; GHOTRA; MCINTOSH; HASSAN, 2015).

Durante o processo criação de um modelo de classificação, é necessário ainda definir um método de avaliação para mensurar e avaliar de forma objetiva o modelo implementado. O processo de avaliação tem como objetivo validar o classificador construído e desta forma, assegurar que o mesmo apresenta um bom desempenho dentro dos padrões aceitáveis, em relação ao evento que ele pretende classificar (JAMES et al., 2013; GHOTRA; MCINTOSH; HASSAN, 2015).

4.4 Métodos de amostragem

Para assegurar a confiabilidade de um modelo, um ponto importante que deve ser considerado ao avaliar-se diferentes algoritmos de ML é a definição do método de amostragem que será utilizado (JAMES et al., 2013). Testar o modelo desenvolvido utilizando o mesmo conjunto de dados de treinamento pode ser enganoso, pois o modelo pode funcionar perfeitamente com o conjunto de dados de treinamento, mas apresentar um desempenho insatisfatório durante a classificação de novos registros (FERRI; HERNÁNDEZ-ORALLO; MODROIU, 2009; JAMES et al., 2013).

Os métodos de amostragem possibilitam a formação de conjuntos de dados distintos para as etapas de treinamento e teste (MITCHELL, 1997; JAMES et al., 2013; WITTEN et al., 2016). A divisão do conjunto de dados tem por objetivo avaliar a capacidade de generalização (classificação de registros desconhecidos) do modelo e atenuar possíveis problemas de *overfitting* e *underfitting* (WITTEN et al., 2016). Existem diversas técnicas de amostragem de dados, entre elas a validação cruzada (CV — *Cross-Validation*), que é a mais recomendada para a atividade de particionamento de um conjunto de dados (MITCHELL,

1997; JAMES et al., 2013). Essa técnica é utilizada para avaliar como os resultados de um algoritmo de ML se comporta ao receber conjuntos de dados independentes (WITTEN et al., 2016). A seguir são apresentados alguns métodos de validação cruzada:

- **Holdout:** Neste método, o conjunto de dados é dividido aleatoriamente em um *dataset* de treinamento e outro de teste, de tamanhos previamente fixados (MITCHELL, 1997). Tipicamente 2/3 dos registros são utilizados para o treinamento e o restante dos dados é usado para testar o modelo, embora outras proporções possam ser utilizadas (KOHAVI et al., 1995; WITTEN et al., 2016).
- **k-Fold Cross-Validation:** Este método consiste em dividir aleatoriamente o conjunto de dados em k segmentos chamados de *folds*. O número de segmentos pode variar conforme a quantidade de registros e a proporção de exemplos em cada classe (KOHAVI et al., 1995). Dos k segmentos gerados, $(k - 1)$ são utilizados para o treinamento do algoritmo de ML e o restante é utilizado para a fase de teste. Este processo é repetido k vezes até que todos os *folds* tenham sido utilizados na etapa de teste (JAMES et al., 2013).
- **Leave-one-out:** Este método, assim como o *k-Fold Cross-Validation*, divide o conjunto de dados em k segmentos. Contudo, neste método o valor de k é definido pelo número de registros existentes no conjunto de dados. Deste modo, $(k - 1)$ registros são utilizados para treinar o classificador e apenas um registro é utilizado para testar o modelo. Esse estimador apresenta um alto custo computacional, uma vez que esse processo é repetido k , para todos os registros do conjunto de dados (JAMES et al., 2013; WITTEN et al., 2016).

4.4.1 Métricas de avaliação de desempenho

A construção de um modelo para análise e classificação da confiabilidade de informação vai além da aplicação de algoritmos de ML sobre um conjunto de dados. Desenvolver um modelo e confrontá-lo com novos dados não é suficiente para determinar se ele apresenta uma boa aderência em relação ao fenômeno que pretende predizer. Para assegurar uma aderência dentro de padrões aceitáveis, é necessário, portanto, calcular e avaliar as métricas de desempenho apresentadas pelo modelo (MITCHELL, 1997; WITTEN et al., 2016).

Na literatura podem ser encontradas inúmeras medidas que são utilizadas individualmente ou em conjunto para realizar a avaliação de modelos de predição (KOHAVI et al., 1995; FAWCETT, 2006; FERRI; HERNÁNDEZ-ORALLO; MODROIU, 2009; WITTEN et al., 2016). Para este estudo, consideramos as seguintes métricas de desempenho: (i) matriz de confusão, (ii) acurácia, (iii) precisão, (iv) sensibilidade, (v) especificidade, (vi) *F-score*, (vii) curva ROC (*Receiver Operating Characteristics Curve*) e (viii) MCC (*Matthews correlation coefficient*).

A matriz de confusão é uma estrutura de classificação que consolida a quantidade de registros realmente pertencentes a cada uma das classes (WITTEN et al., 2016). Conforme exemplificado na Figura 7, essa estrutura apresenta os seguintes valores:

- **Verdadeiro Positivo (TP – *True Positive*)** que representa o total de registros preditos como pertencentes à classe positiva e que realmente pertencem a essa classe;
- **Falso Positivo (FP – *False Positive*)** que descreve o total de registros preditos como pertencentes à classe positiva, mas que na verdade pertencem à classe negativa;
- **Verdadeiro Negativo (TN – *True Negative*)** que demonstra o total de registros preditos como pertencentes à classe negativa e que realmente pertencem à essa classe;
- **Falso Negativo (FN – *False Negative*)** que apresenta o total de registros preditos como pertencentes à classe negativa mas que na verdade pertencem à classe positiva,

Figura 7 – Exemplo de matriz de confusão.

Classe Real	Classe Predit	
	Positiva	Negativa
Positiva	TP	FN
Negativa	FP	TN

Fonte: Elaborado pelo autor.

Utilizando os resultados de uma matriz de confusão, é possível calcular outros indicadores como, por exemplo, a acurácia que é utilizada para descrever a capacidade apresentada pelo modelo em determinar rótulos de classe corretos para registros novos. Seu valor pode ser obtido conforme descrito na Equação 3. É importante ressaltar que

essa medida não é um indicador confiável de bom desempenho para conjuntos de dados não-balanceados, em que a proporção entre classes não é a mesma.

$$\text{Acurácia} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3)$$

A partir das informações obtidas em uma matriz de confusão, é possível ainda, derivar as medidas de precisão, sensibilidade e especificidade, conforme descrito nas equações 4, 5 e 6, respectivamente.

$$\text{Precisão} = \frac{TP}{(TP + FP)} \quad (4)$$

A precisão pode ser descrita como a razão entre o número de registros classificados corretamente para a classe positiva (TP), dividido pelo total de registros classificados para esta classe ($TP + FP$).

$$\text{Sensibilidade} = \frac{(TP)}{(TP + FN)} \quad (5)$$

A sensibilidade, também conhecida como revocação (*recall*) ou taxa de verdadeiros positivos, descreve a razão entre a quantidade de registros corretamente classificados para a classe positiva (TP), e o total de registros que pertencem a esta classe ($TP + FN$). Por sua vez, a especificidade ou taxa de verdadeiros negativos é utilizada para descrever a razão entre a quantidade de registros corretamente classificados para a classe negativa (TN) e o total de registros que pertencem a esta classe ($TN + FP$).

$$\text{Especificidade} = \frac{(TN)}{(TN + FP)} \quad (6)$$

Os valores de precisão, sensibilidade e especificidade podem ser utilizados para calcular outras medidas, tais como F-score e curva ROC. A métrica *F-score* é utilizada para representar a média ponderada dos valores de precisão e sensibilidade. Seu valor é obtido de mediante a utilização da Equação 7.

$$\text{F-score} = 2 \times \left(\frac{(\text{Precisão} \times \text{Sensibilidade})}{(\text{Precisão} + \text{Sensibilidade})} \right) \quad (7)$$

Para avaliar a habilidade que um modelo apresenta em classificar corretamente um novo registro, é possível utilizar ainda a Curva ROC. Esse recurso consiste em um gráfico bidimensional, utilizado para descrever de forma visual a probabilidade de verdadeiros

positivos (*sensibilidade*) em relação aos falsos positivos (*especificidade*) (FAWCETT, 2006; BRINK; RICHARDS; FETHEROLF, 2017). Por convenção, a especificidade é representada no eixo das abscissas (eixo horizontal) e a sensibilidade é representada no eixo das ordenadas (eixo vertical), variando ambas de 0 a 1 (0–100%). A diagonal que corta o espaço ROC ao meio é conhecido como a linha de zero discriminação, onde permanecem os resultados de decisões aleatórias. Logo, os pontos acima desta linha indicam um desempenho superior às decisões aleatórias, enquanto pontos abaixo da linha indicam desempenho inferior (BRINK; RICHARDS; FETHEROLF, 2017; GHOTRA; MCINTOSH; HASSAN, 2015). A Figura 8 ilustra um exemplo de uma estrutura de gráfico de curva ROC.

Figura 8 – Exemplo de um gráfico de Curva ROC

Fonte: Elaborado pelo autor.

A partir de um gráfico ROC é possível calcular uma medida geral de qualidade, a área sob a curva (AUC — *Area Under the Curve*), também conhecida por *c-statistic* (HAND, 1997; FAWCETT, 2006). A AUC é a fração da área total que se situa sob a curva ROC e apresenta valores que variam de 0,5 no mínimo, quando o modelo não consegue classificar nenhum caso, a 1 no máximo, quando ocorre 100% de sensibilidade e 100% de especificidade (ZHU et al., 2010; GHOTRA; MCINTOSH; HASSAN, 2015; HOSMER; LEMESHOW; STURDIVANT, 2013). A AUC é vista com uma medida da qualidade apresentada pelo classificador, pois, quanto maior a área, melhor o desempenho do classifi-

cador(HOSMER; LEMESHOW; STURDIVANT, 2013). O Quadro 3, descreve um sistema para avaliar por meio da AUC o poder de classificação apresentado por um modelo.

Quadro 3 – Classificação de desempenho do modelo dado pela AUC.

Valor AUC	Poder de Classificação
$AUC = 0,5$	Não há
$0,7 \leq AUC < 0,8$	Aceitável
$0,8 \leq AUC < 0,9$	Muito bom
$AUC \geq 0,9$	Excelente

Fonte: Ferraz e Torres (2004), adaptado pelo autor.

A medida MCC (*Matthews correlation coefficient*) é utilizada para descrever o coeficiente de correlação entre as classes dependentes, que é utilizada como uma medida de qualidade. Seu valor varia de -1 a 1, onde valores mais próximos de -1 são indicadores de um modelo de predição ruim, valores iguais a 0 indicam que o modelo de predição é totalmente randômico e valores mais próximos de 1 são indicadores de um modelo de predição com bom desempenho. Seu valor é obtido de mediante a utilização da Equação 8.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

O uso das métricas de desempenho apresentadas nesta seção é extremamente importante para esta pesquisa, pois, com a análise dos resultados obtidos a partir dessas métricas é possível mensurar e avaliar de forma objetiva a eficiência apresentada pelos algoritmos utilizados na construção dos modelos de classificação da confiabilidade do conjunto de dados.

5 Trabalhos Relacionados

Neste capítulo apresentaremos os principais trabalhos relacionados ao tema desta dissertação. A Seção 5.1 comenta os trabalhos encontrados na literatura que abordam a temática “o uso de *crowdsourcing* em problemas de cidades inteligentes”. Na Seção 5.2 são apresentados trabalhos que discutem a importância da utilização de métodos e técnicas para avaliar a qualidade e garantir a confiabilidade dos dados obtidos com o uso de *crowdsourcing*.

5.1 Uso de *crowdsourcing* em problemas de cidades inteligentes

O uso de técnicas de *crowdsourcing* em problemas relacionados a cidades inteligentes já foi investigado por outros estudos. Foi realizado um levantamento do estado da arte considerando os trabalhos correlatos aos temas: (i) cidades inteligentes, com ênfase na utilização de técnicas de *crowdsourcing*, (ii) sistemas inteligentes de transporte e (iii) sistemas de informações ao usuário. Realizou-se também um levantamento sobre os principais métodos e abordagens propostos na literatura para a análise de confiabilidade de dados obtidos por *crowdsourcing*.

Tabela 2 – Estudos que utilizaram técnicas de *crowdsourcing*.

Autor(es)	Ano	Seguimento do estudo	Análise a confiabilidade dos dados	Técnicas/Abordagens de análise de confiabilidade
STERNBERG; LANTZ	2018	segurança pública	Não	N/A
CEROTTI et al.	2016	mobilidade urbana	Não	N/A
OLIVEIRA et al.	2016	saúde	Não	N/A
MISRA et al.	2014	mobilidade urbana	Sim	<i>Crowdsourcing</i>
MIRRI et al.	2014	acessibilidade	Não	N/A
ZHOU; ZHENG; LI	2014	mobilidade urbana	Não	N/A
HOAR	2010	mobilidade urbana	Não	N/A
FURTADO et al.	2010	segurança pública	Não	N/A

Os trabalhos encontrados apresentam os resultados referentes à utilização das técnicas de *crowdsourcing* em diversos segmentos relacionados a cidades inteligentes. Os trabalhos incluem resultados sobre: (i) mobilidade urbana, (ii) saúde, (iii) acessibilidade e (iv) segurança pública. A Tabela 2 destaca os principais trabalhos encontrados na literatura

que utilizaram técnicas de *crowdsourcing* como mecanismo de captura de informações complementares.

Sternberg e Lantz (2018) apresentam um aplicativo para dispositivos móveis que permite que seus usuários contribuam com informações referente aos crimes de trânsito (roubo de veículos, roubo de combustível, abuso de autoridade policial, etc.) ocorridos na Suécia. Conforme expõem os autores, existem poucos dados sobre os crimes de trânsito, e a falta desses dados inviabiliza o processo de tomada de decisão e a criação de políticas públicas que atendam as necessidades da população. O trabalho realizado pelos autores descreve como o aplicativo desenvolvido pode contribuir para a obtenção de novas informações que podem ser utilizadas para auxiliar em uma possível análise de fatores que contribuem para ocorrência de tais crimes.

O trabalho proposto por Cerotti et al. (2016) apresenta o desenvolvimento da ferramenta MoCSACO, que utiliza métodos de Markov (modelos matemáticos utilizados para realizar o reconhecimento de padrões em conjuntos de dados com variáveis aleatórias) em conjunto com as técnicas de *crowdsourcing* para realizar a monitoração e planejamento das melhores rotas de trânsito. Segundo os autores a aplicação desenvolvida pode ser utilizada para disponibilizar informações para os seus usuários sobre as condições de trânsito de uma determinada área em tempo real. Os autores expõem que os dados utilizados pelo MoCSACO são coletados utilizando técnicas de *crowdsourcing*. A captura de informações é realizada a partir dos dispositivos móveis dos próprios motoristas que utilizam o aplicativo. Os dados coletados são processados utilizando um conjunto de agentes markovianos que são capazes de realizar a análise interativa dessas informações e posteriormente utilizam os padrões descobertos para sugerir novas rotas para os usuários.

Oliveira et al. (2016) apresentam uma proposta de jogo interativo que utiliza *crowdsourcing* associado a técnicas de gamificação para conscientizar e solicitar o engajamento da população no combate ao mosquito *Aedes aegypti*. Segundo os autores, o jogo desenvolvido utiliza um sistema de pontuação e classificação que tem como objetivo incentivar o usuário a contribuir com a realização de várias atividades como, por exemplo, denunciar possíveis focos de *Aedes aegypti*, informar sobre os casos confirmados de doenças (Zika, Chikungunya, Dengue, Febre Amarela, etc.) causadas pelos mosquitos e agendar visita a imóvel.

O estudo realizado por Misra et al. (2014) propõe a utilização de *crowdsourcing* como uma possível ferramenta de coleta de dados, para ser utilizada no planejamento

e controle dos serviços de transporte, bem como para solicitar *feedback* dos usuários sobre qualidade desses serviços. De acordo com os autores a utilização de técnicas de *crowdsourcing* possibilita a obtenção de diversos tipos de dados que normalmente não poderiam ser capturados por meios das abordagens tradicionais. Contudo, Misra et al. (2014) explicam que os dados coletados com *crowdsourcing* podem apresentar informações não confiáveis. Sendo assim, os autores recomendam que seja realizada uma análise da qualidade dessas informações antes de serem integradas a base de dados de um SIU. No trabalho desenvolvido é proposta ainda a utilização da própria abordagem de *crowdsourcing* para realizar a análise da qualidade das informações.

O trabalho relatado por Mirri et al. (2014) descreve duas aplicações: mPASS e WhenMyBus. Esses aplicativos utilizam técnicas de *crowdsourcing* para obter dados sobre a qualidade dos serviços e estruturas (rampas, piso tátil, barras de apoio, etc.) utilizados por pessoas com algum tipo de deficiência e têm como principal objetivo a disponibilização de informações em tempo real, sobre as condições dos recursos de acessibilidade existentes. Os autores explicam que os sistemas são baseados no OpenStreetMap¹ e realiza a combinação das informações obtidas das bases de dados de várias prestadoras de serviços públicos com os dados disponibilizados voluntariamente pelos próprios usuários a partir de seus dispositivos móveis. Além disso, segundo os autores os aplicativos combinam as informações coletadas com os dados informados no perfil do usuário para disponibilizar os melhores trajetos a serem realizados de acordo com as necessidades de cada pessoa.

A pesquisa realizada por Zhou, Zheng e Li (2012) descreve o desenvolvimento de um sistema de previsão de tempo de chegada de ônibus, que utiliza um conjunto de métodos estatísticos com as informações disponibilizadas voluntariamente pelos passageiros por meio dos seus dispositivos móveis. Conforme expõem os autores, essa aplicação não depende das informações das prestadoras de serviços de transporte. As predições de tempo de chegada é realizada exclusivamente a partir dos dados obtidos dos dispositivos móveis dos próprios passageiros. Zhou, Zheng e Li (2012) explicam ainda que o sistema desenvolvido não utiliza o recurso de GPS dos aparelhos dos usuários para coletar a localização do ônibus. De acordo com os autores, a localização do veículo é calculada utilizando a potência de sinal apresentada entre as torres de celular da região e os dispositivos móveis dos usuários. Deste modo, os dados coletados a partir dos celulares dos passageiros são transmitidos

¹ Projeto de mapeamento colaborativo disponível em: <https://www.openstreetmap.org/>

ao servidor, onde posteriormente são processados e as informações geradas são utilizadas para estimar o horário de chegada no próximo ponto de parada.

O trabalho realizado por Hoar (2010) apresenta os resultados obtidos durante o desenvolvimento de aplicações para dispositivos móveis que tem por objetivo disponibilizar informações para os usuários e gestores de transporte público, como, por exemplo, rotas, mapas, horários de parada entre outras informações. A aplicação desenvolvida apresenta uma funcionalidade que permite que os passageiros compartilhem informações em tempo real sobre a localização atual dos veículos e se eles estão cumprindo o horário anteriormente determinado. Os dados coletados são utilizados para realizar a criação de relatórios contendo informações relevantes sobre os serviços de transporte. Como resultado, o autor expõe que a partir dos relatórios gerados, é possível fornecer uma melhor visibilidade dos problemas existentes para os gestores de serviços de transportes públicos, contribuindo deste modo, para aumentar a qualidade desses serviços. Além disso, as informações disponibilizadas em tempo real para os passageiros por meio de seus dispositivos, possibilitam que os mesmos realizem um planejamento melhor de suas viagens.

A pesquisa realizada por Furtado et al. (2010) descreve o desenvolvimento do sistema WikiCrimes, criado com o objetivo de mapear a ocorrência de crimes colaborativamente, permitindo que os cidadãos visualizem a densidade de crimes por região, bem como os tipos de ocorrência registrados. A aplicação desenvolvida pelos autores estimula, ainda, o envio de denúncias às autoridades competentes. Como resultado, os autores destacam a criação de uma plataforma que disponibiliza diversas informações dinâmicas, desde estatísticas criminais sobre uma determinada região, até mapas temáticos e diversos indicadores relacionados à criminalidade que podem ser utilizados pelos órgãos públicos para criar ações de melhorias relacionadas à segurança pública.

5.2 Confiabilidade de *crowdsourcing*

Conforme apresentado na seção anterior, existem diversos estudos que abordam a utilização de *crowdsourcing* como recurso inovador para solucionar os problemas relacionados às estruturas de SC. Tais trabalhos apresentam e descrevem como o uso de dados obtidos com *crowdsourcing* pode ajudar na obtenção de informações que normalmente não poderiam ser capturadas por outras abordagens tradicionais.

No contexto de mobilidade urbana, os trabalhos realizados por Hoar (2010), Misra et al. (2014), Zhou, Zheng e Li (2012) e Cerotti et al. (2016) são os estudos que mais se assemelham ao presente trabalho. Tais trabalhos centraram no problema de construir SIU eficientes e confiáveis, e em como permitir uma melhor interação entre usuários e gestores de serviços de transporte público. Os dados necessários para a criação destes sistemas são fornecidos pelos próprios usuários, por meio dos seus respectivos dispositivos móveis, o que possibilita a atualização constante dos dados. Contudo, nenhum desses trabalhos, com exceção de (MISRA et al., 2014), preocupou-se em analisar e assegurar a qualidade e confiabilidade dos dados fornecidos pelos usuários. É importante destacar ainda que o trabalho desenvolvido por Misra et al. (2014) não utiliza uma abordagem que realiza tal análise de forma automatizada. A validação da qualidade dos dados nesse estudo também foi feita pelos próprios usuários, com o auxílio de técnicas de *crowdsourcing*.

Assim como ocorre com outras abordagens, os dados obtidos via *crowdsourcing* também podem apresentar informações não confiáveis, que comprometem a confiabilidade dos dados coletados. A qualidade dos dados coletados utilizando técnicas de *crowdsourcing* está diretamente ligada à motivação e à interação de seus colaboradores. Além da disponibilidade dos usuários para colaborar com informações, existem outros fatores como, por exemplo, dificuldade de interação com o aplicativo e qualidade dos recursos tecnológicos dos dispositivos móveis (conexão 3G/4G, GPS, etc.), entre outros, que podem contribuir para que os dados coletados apresentem informações não confiáveis.

Diversos autores (MASHHADI; CAPRA, 2011; ALLAHBAKHSI et al., 2013; JAGABATHULA; SUBRAMANIAN; VENKATARAMAN, 2014; MOUSA et al., 2015; DANIEL et al., 2018) estudaram a necessidade da definição e adoção de métodos e técnicas para avaliar a qualidade e garantir a confiabilidade dos dados obtidos com o uso de *crowdsourcing*. Entretanto, embora a confiabilidade dos dados coletados com uso de *crowdsourcing* seja um fator crucial, ainda não existe uma abordagem específica que possibilite avaliar de forma eficiente a qualidade dessas informações.

Nota-se que nenhum dos trabalhos apresentados nesta seção investigou o uso de técnicas de aprendizado de máquina para a análise de confiabilidade de dados obtidos via *crowdsourcing*. No próximo capítulo, serão apresentados os modelos de análise de confiabilidade de dados que permitirá mitigar as incertezas introduzidas pelo uso de *crowdsourcing*.

6 Análise de confiabilidade de dados de *crowdsourcing* com Aprendizado de Máquina

Este capítulo apresenta os modelos construídos e utilizados no processo de análise de confiabilidade das informações sobre a localização dos ônibus da rede de transporte público da cidade de São Paulo, obtidas mediante o uso de *crowdsourcing*. Na Seção 6.1 é apresentada a descrição do modelo proposto e desenvolvido por essa pesquisa. A Seção 6.2, descreve o conjunto de dados utilizados. Já na Seção 6.3 é apresentado o processo, bem como, a estrutura utilizada durante a coleta de dados. A Seção 6.4 expõe os procedimentos envolvidos no processo de análise e seleção dos registros que compõem o conjunto de dados final utilizado. Por fim, a Seção 6.5, descreve detalhadamente como cada um dos modelos implementados foi avaliado.

6.1 Descrição do modelo de análise da confiabilidade

Uma preocupação recorrente no desenvolvimento de cidades inteligentes é a otimização dos sistemas de transporte público. O sistema de transporte público é um recurso que afeta significativamente a qualidade de vida da população de grandes metrópoles. Em particular, uma preocupação importante diz respeito à qualidade dos serviços prestados, especialmente quando nos referimos ao modal ônibus como meio de transporte coletivo. As prestadoras de serviços têm dificuldades para estabelecer e cumprir horários de itinerários das linhas de ônibus. Tais problemas, associados à falta de um sistema confiável e previsível, levam o usuário a não optar pela utilização desses serviços, o que agrava problemas urbanos sociais e ambientais.

Para tentar reverter esse quadro, as iniciativas em SC propõem a utilização de dados coletados por sistemas distribuídos, pertencentes às diversas estruturas tecnológicas como, por exemplo, os ITS. Entretanto, diversos fatores podem contribuir para que esses dados sejam insuficientes. Os dados obtidos com *crowdsourcing* são eficazes neste cenário (HOAR, 2010; MISRA et al., 2014; ZHOU; ZHENG; LI, 2012; CEROTTI et al., 2016) e apresentam um grande potencial como abordagem de levantamento de informações no contexto de transportes públicos, pois, permitem a obtenção de novas informações em tempo real (MISRA et al., 2014). Contudo, os dados obtidos com *crowdsourcing*

indica um dado incorreto reportado pelo usuário. Em casos assim, o modelo construído deverá classificar este tipo de informação como uma contribuição de *crowdsourcing* “não confiável”.

Figura 10 – Exemplo de situação onde o dado informado pelo usuário é confiável.

Fonte: Elaborado pelo autor.

Já a Figura 10 apresenta um exemplo em que é possível observar as contribuições que as informações obtidas por meio de técnicas de *crowdsourcing* podem proporcionar. Neste cenário, de acordo com os dados apresentados pela Figura 10a durante a realização do trajeto de um ônibus, no intervalo de tempo entre as 00h 06min 04s e 00h 09min 11s, por algum motivo não existe nenhum tipo de registro coletado pelo sistema AVL instalado no veículo. No entanto, durante esse mesmo intervalo de tempo, precisamente às 00h 07min 38s, foi identificada a ocorrência de uma contribuição, realizada por um usuário do aplicativo Coletivo.

A Figura 10b ilustra ainda a localização dos registros coletados, tornando possível identificar que a localização informada pelo usuário corresponde ao trajeto realizado pelo veículo. Nestes casos, o modelo desenvolvido deverá classificar esse tipo registro com uma contribuição de *crowdsourcing* confiável.

6.2 Conjunto de dados

Os algoritmos de ML utilizados para a construção dos modelos de análise de dados de *crowdsourcing* para o transporte público utiliza um *dataset* que contém informações de duas bases: a da *SPtrans*, que contém os dados disponibilizados pela empresa São Paulo Transporte S. A. (SPTrans¹) e a *base de dados crowdsourcing*, disponibilizada pela *startup* Scipopulis², especializada em mobilidade urbana.

Tabela 3 – Descrição do conjunto de dados da SPTrans.

Nome do campos	Descrição
<i>cd_evento_avl_movto</i>	Código sequencial identificador do evento
<i>cd_linha</i>	Código identificador da linha em operação
<i>dt_movto</i>	Data da gravação em banco de dados do evento gerado no AVL
<i>nr_identificador</i>	Código identificador do AVL
<i>nr_evento_linha</i>	Grupo de indicadores relacionados ao evento
<i>nr_ponto</i>	Código do ponto notável
<i>nr_velocidade</i>	Velocidade instantânea
<i>nr_voltagem</i>	Tensão de alimentação
<i>nr_temperatura_interna</i>	Temperatura do processador
<i>nr_evento_terminal_dado</i>	Código do evento relacionado no terminal de dados (botões)
<i>nr_evento_es_1</i>	Grupo de indicadores relacionados ao evento
<i>nr_latitude_grau</i>	Latitude da geolocalização do veículo
<i>nr_longitude_grau</i>	Longitude da geolocalização do veículo
<i>nr_indiceregistro</i>	Índice de geração do evento no AVL
<i>dt_avl</i>	Data da geração do evento no AVL
<i>nr_distancia</i>	Distância em metros do evento com relação ao evento anterior do mesmo AVL
<i>cd_tipo_veiculo_geo</i>	Código para identificação no software de mapeamento
<i>cd_avl_conexao</i>	Código interno utilizado para identificar qual a conexão utilizada para transmissão do evento
<i>cd_prefixo</i>	Prefixo do veículo

¹ Empresa responsável pela gestão do sistema de transporte da cidade de São Paulo.: <http://www.sptrans.com.br/>

² *Startup* de inovação para cidades inteligentes e dedicada à mobilidade urbana, financiada pela FAPESP: <http://scipopulis.com/>

A primeira base, conforme descrito na Tabela 3, contém as informações dos registros referentes aos trajetos realizados pelos ônibus da rede de transporte público da cidade São Paulo. Esses dados foram coletados pelos sistemas AVL instalados em cada veículo. A SPTrans disponibiliza os dados coletados em tempo real para os desenvolvedores, acrescidos de informações como: identificação da linha, identificação e localização dos veículos e horários em que foram realizados os registros, entre outras. A base de dados disponibilizada contém mais de 800 milhões de registros, totalizando 720 GB de dados.

Tabela 4 – Descrição do conjunto de dados de *crowdsourcing*.

Nome do campos	Descrição
<i>bus_id</i>	Código sequencial identificador do veículo
<i>dt_mvto</i>	Data da gravação do evento realizado pelo usuário
<i>lat</i>	Latitude da geolocalização do usuário
<i>lng</i>	Longitude da geolocalização do usuário
<i>route</i>	Código identificador da linha do ônibus
<i>stop_id</i>	Identificador do ponto de parada informado pelo usuário (relacionado ao GTFS da cidade)

Já a segunda base (descrita pela Tabela 4) contém os dados privados fornecidos pela *startup* Scipopulis, com informações disponibilizadas voluntariamente pelos usuários do aplicativo Coletivo, descrito na Seção 3.3. Esse aplicativo possui uma funcionalidade que permite que usuários informem se um determinado ônibus já passou enquanto o usuário aguarda seu ônibus. Quando o passageiro informa que o ônibus chegou no ponto de parada, o sistema registra informações como (i) horário (data e hora atual em que registro foi realizado), (ii) localização (latitude e longitude obtida pelo GPS do usuário), (iii) identificador do ponto de parada e (iv) número da linha do ônibus. Essa base contém 16.894 (dezesesseis mil e oitocentos e noventa e quatro) registros, totalizando 3 GB de dados. A Figura 11 apresenta uma visão geral da distribuição dos dados coletados por *crowdsourcing* na cidade de São Paulo.

Figura 11 – Mapa de distribuição de localizações das contribuições coletivas.

Fonte: Elaborado pelo autor.

6.3 Coleta de Dados

Os dados coletados são referentes aos trajetos realizados por 992 linhas de ônibus da cidade de São Paulo, durante o período de 1º de novembro de 2016 à 31 de maio de 2017. A Figura 12 apresenta, de forma esquemática, como os dados foram coletados.

Figura 12 – Coleta de dados de posicionamento dos ônibus.

Fonte: Elaborado pelo autor.

É importante ressaltar que nem todos os registros contidos nas bases foram utilizados. O conjunto de dados final utilizado nesse estudo é composto pelo relacionamento entre as informações coletadas com *crowdsourcing* e os dados fornecidos pela SPTrans, ou seja,

somente os registros capturados pelos sistemas AVL que apresentam algum relacionamento com as informações disponibilizadas pelos usuários foram considerados.

Além disso, os dados coletados tanto pelos sistemas AVL quanto pelo uso de *crowdsourcing*, apresentam informações ausentes ou erradas, devido às falhas de equipamentos, problemas de comunicação, e outros fatores que reduzem a qualidade dos dados (PAPAGIANNAKI et al., 2003; WANG; SHI, 2013; MACEDO, 2015). Diante da quantidade de informações coletadas, foi necessário realizar um pré-processamento desses dados para selecionar as informações necessárias, formatá-las, e remover as possíveis inconsistências contidas nas bases de dados.

6.4 Análise dos dados

Conforme exemplificado na Figura 13 os dados coletados e utilizados nesse trabalho passaram pelos seguintes processos antes de serem utilizados: (i) *consolidação dos dados* em uma única base dos dados referentes aos trajetos realizados pelos ônibus; (ii) *estruturação e normalização* para remoção de dados duplicados ou que apresentam falta de informações, (iii) *classificação dos dados*.

Figura 13 – Etapas do processo de análise e tratamento dos dados.

Fonte: Elaborado pelo autor.

6.4.1 Consolidação dos dados

Para criar um conjunto de dados adequado para os algoritmos de ML implementados, inicialmente foi realizada a criação de uma única estrutura contendo todas as informações

referentes aos trajetos realizados pelos ônibus. Durante essa etapa, foi realizada também uma análise dos conjuntos de dados para verificar quais as variáveis (atributos) estavam dentro do escopo de interesse do trabalho. Para melhorar a qualidade das informações, o conjunto de dados foi enriquecido com o auxílio das informações presentes no GTFS (*General Transit Feed Specification*).

O GTFS é um formato padrão de dados de transporte público (que contém informações sobre a programação dos itinerários e os seus dados geográficos), que foi desenvolvido em uma parceria entre Google e a TriMet³, a agência de transportes de Portland, Oregon. O GTFS fornece dados estáticos sobre a rede de transportes de uma cidade. A partir desses dados é possível saber os trajetos dos ônibus, a localização dos pontos de parada e a frequência de passagem de cada linha. Essas informações são disponibilizadas em diversos arquivos no formato CSV, compactados em um único arquivo com formato ZIP (PONS; MONTEIRO; CARDOSO, 2015).

Para esse estudo, foram consideradas as seguintes informações dos arquivos GTFS: *(i)* identificador dos pontos de parada e *(ii)* localização dos pontos de parada obtidas por meio das coordenadas de *(a)* latitude e *(b)* longitude. Com base nessas informações, foram criadas ainda as variáveis derivadas *(a)* distância da localização do usuário, *(b)* última distância do ônibus, e *(c)* última velocidade média. Tais variáveis contém informações relevantes para os modelos de análise de confiabilidade desenvolvidos.

A distância da localização do usuário, foi obtida mediante o cálculo da distância apresentada entre a localização capturada a partir dos dispositivos móveis dos passageiros e as coordenadas de latitude e longitude do ponto de parada atualmente disponível no formato GTFS. O relacionamento entre essas informações foi realizado por meio do número do *(i)* identificador do ponto de parada informado pelo usuário e o *(ii)* identificador do ponto de parada correspondente no GTFS.

Já a última distância dos ônibus, foi obtida mediante o cálculo da distância apresentada entre a última localização capturada a partir do sistema AVL instalado no veículo e as coordenadas de latitude e longitude do ponto de parada atualmente disponível no GTFS correspondente ao ponto de parada informado pelo usuário. Por fim, as últimas velocidades médias foram calculadas considerando o tempo de deslocamento apresentado

³ Agência pública responsável pela gestão do sistema de transporte da cidade de Portland (<http://trimet.org>)

pelo ônibus em relação à distância apresentada entre o ponto de parada informado pelo usuário e a última localização capturada pelo sistema AVL instalado no veículo.

6.4.2 Estruturação e normalização dos dados

O processo de normalização de dados é realizado para que se possa extrair os dados de interesse para a pesquisa (WANG; SHI, 2013; MACEDO, 2015). Esse processo permite a remoção de informações que, por algum motivo apresentem inconformidades e/ou anomalias (duplicidades, falta de informação, etc.), ou que não sejam relevantes para a análise em questão. Essa etapa permite determinar os dados que serão considerados durante as atividades de análise de confiabilidade. Sendo assim, essa etapa influencia diretamente na qualidade do conjunto de dados utilizado (PAPAGIANNAKI et al., 2003).

Essa etapa encontrou vários problemas com os dados analisados. O envio de informações para o servidor em que são armazenados os dados coletados, por exemplo, não obedece um intervalo regular de tempo e, devido aos problemas de conexão inerentes às tecnologias utilizadas, muitos dos registros analisados apresentavam duplicidades. Nestes casos, foi mantido apenas o registro que apresentava o menor horário, ou seja, o primeiro registro recebido pelo servidor.

6.4.3 Classificação dos dados

Conforme mencionado na Seção 4.1, o método de aprendizado supervisionado utiliza um *dataset* onde as classes (ou *rótulos*) já são conhecidas. Portanto, após a etapa de estruturação e normalização, foi realizada a classificação dos dados. Nesta fase, cada registro do conjunto de dados recebeu uma classificação (confiável ou não confiável), de acordo com a qualidade das informações apresentadas. A classificação foi atribuída considerando os atributos referentes aos registros efetuados pelos usuários em relação aos trajetos realizados pelos veículos da rede de transporte coletivo da cidade de São Paulo. Essa classificação foi realizada seguindo um protocolo de validação, onde foram consideradas:

- **Distância do usuário em relação ao ponto de ônibus** — (a) inversamente proporcional à probabilidade do dado ser confiável, se inferior a 300 metros ou (b) não confiável, caso contrário;
- **Distâncias anteriores e posteriores do veículo em relação ao ponto de parada informado pelo usuário** — (a) inversamente proporcional à probabilidade do dado ser confiável, se inferior a 10 quilômetros ou (b) não confiável, caso contrário;
- **Velocidade média entre as localizações anteriores e posteriores do veículo em relação à localização do ponto de ônibus informada pelo usuário** — (a) diretamente proporcional à probabilidade do dado ser confiável, se superior 10 km/h e inferior a 80 km/h ou (b) não confiável, caso contrário;
- **Tempos e distâncias dos registros realizados pelos usuários em relação à localização e aos tempos anteriores e posteriores do AVL** — (a) inversamente proporcional à probabilidade do dado ser confiável, se a diferença de tempo entre o registro capturado pelo sistema AVL em relação ao registro informado pelo usuário for inferior a 15 minutos ou (b) não confiável, se superior 15 minutos, e (c) diretamente proporcional à probabilidade do dado não ser confiável quando o tempo e a distância entre os registros capturados pelo equipamento AVL em relação ao registro informado pelo usuário forem inversamente proporcionais.

Por fim, após a execução das etapas anteriores, chegamos ao conjunto de dados final, contendo 13.493 (treze mil e quatrocentos e noventa e três) registros. O conjunto de dados selecionado considera as seguintes informações: *(i)* identificadores das linhas, *(ii)* dos veículos e *(iii)* dos pontos de paradas; *(iv)* tipo de trajeto realizado (ida = 1 e volta = 0); *(v)* distância do usuário em relação ao ponto de parada, obtida por meio das coordenadas de *(a)* latitude e *(b)* longitude capturadas a partir dos dispositivos móveis dos passageiros; *(vi)* os tempos em que foram realizados os registros pelos usuários, formados por *(a)* data e *(b)* hora; *(vii)* distância da localização anterior do veículo em relação ao ponto de parada, obtida por meio das coordenadas de *(a)* latitude e *(b)* longitude, capturadas a partir dos sistemas AVL instalados nos veículos; *(viii)* os tempos em que foram capturadas as informações do sistema AVL, formados por *(a)* data e *(b)* hora; *(ix)* velocidade média necessária para o deslocamento do veículo entre a localização anterior registrada pelo sistema AVL e a nova localização informada pelo usuário; e *(x)* classificação da qualidade

da informação (confiável = 1 e não confiável = 0). Alguns registros de dados analisados são apresentados na Tabela 5.

Tabela 5 – Exemplos de registros dos dados de *crowdsourcing* analisados.

Nº Linha	Nº ônibus	Ponto parada	Trajetos ⁽¹⁾	Distância usuário (m)	Tempo usuário	Distância anterior (m)	Tempo anterior	Velocidade anterior (km/h)	Classificação ⁽²⁾
5142-10	51696	140017022	0	102,79	11/01/16 11:52	361,13	11/01/16 11:51	35,14	1
719R-10	81085	120015876	1	169	11/01/16 12:21	788	11/01/16 12:20	91,51	0
5108-10	53511	140017022	0	104,64	11/01/16 12:38	524,97	11/01/16 12:36	24,23	1
177H-10	21355	120010351	1	17,03	11/01/16 03:24	579,88	11/01/16 03:23	31,16	1
8700-10	82418	630015010	1	30,55	11/01/16 07:30	686,73	11/01/16 07:29	48,48	1
677A-10	73796	3407115	1	42,66	11/01/16 08:10	479,72	11/01/16 08:09	52,33	1
5154-10	76033	530015285	0	4,28	11/01/16 09:44	181,32	11/01/16 09:43	16,32	1
5154-10	76858	3407120	0	33,78	11/01/16 10:21	238,29	11/01/16 10:20	16,82	1
675A-10	77835	790015062	0	13,61	11/01/16 10:27	415,14	11/01/16 10:27	498,16	0
675A-10	77835	790015062	0	11,93	11/01/16 10:27	415,14	11/01/16 10:27	31,14	1

1) 0 = Ida 1 = Volta

2) 0 = Não Confiável 1 = Confiável

6.5 Comparação entre os modelos de análise de confiabilidade

A confiabilidade dos dados foi analisada usando 5 algoritmos de aprendizado de máquina diferentes: (i) Regressão logística, (ii) *K-Nearest Neighbour*, (iii) Árvore de Decisão, (iv) *Gaussian Naive Bayes*, e (v) Máquinas de Vetores Suporte. Para a implementação dos algoritmos foi utilizada a biblioteca *Scikit-Learn*⁴ (PEDREGOSA et al., 2011). Para cada algoritmo implementado, a confiabilidade do conjunto de dados foi analisada considerando-se 4 variáveis: uma variável resposta (y) e 3 variáveis independentes (X), a saber:

- **Classificação:** variável resposta, indica a qualidade da informação dos registros analisados;
- **Distância do usuário:** variável independente que indica a distância entre a localização do usuário e o ponto de parada informado pelo passageiro;
- **Distância anterior:** variável independente que indica a distância entre o ponto de parada informado pelo usuário e a localização anterior do veículo;

⁴ *Scikit-Learn* é uma biblioteca de código *open source* que possui diversos algoritmos de aprendizado de máquina implementados na linguagem de programação Python.

- **Velocidade:** variável independente que indica a velocidade média necessária para o deslocamento entre a localização anterior do veículo e o ponto de parada informado pelo usuário.

Cada um dos algoritmos implementados por esse estudo foi comparado considerando o desempenho apresentado pelos seus classificadores quanto a sua capacidade de identificar de forma eficiente registros que, por algum motivo, apresentam informações não confiáveis (ausentes ou erradas).

6.5.1 Seleção dos método de amostragem

Neste trabalho, a divisão do conjunto de dados foi realizada mediante a utilização do método *K-fold cross-validation*, descrito na Seção 4.4. De todo o conjunto de dados (com 13.493 registros), 70% (9.445 dos registros) da amostra foram separados para serem utilizados na etapa de treinamento dos modelos. Sendo assim, os 30% (4.048 registros) restante da amostra foram utilizados para testar os modelos gerados.

Figura 14 – Divisão do conjunto de dados utilizando *K-fold cross-validation*.

Fonte: Elaborado pelo autor.

Conforme ilustrado na Figura 14, após a divisão dos dados em 2 amostras, o *dataset* destinado à etapa de treinamento foi submetido ao método *K-fold cross-validation*, que consiste em dividir o conjunto de dados aleatoriamente em k segmentos, chamados de *folds*.

Cada um dos algoritmos implementados foi treinado com a iteração de 10-*folds*, conforme recomendações de diversos autores (KOHAVI et al., 1995; WITTEN et al., 2016).

Após cada algoritmo implementado ser devidamente treinado, o mesmo foi apresentado ao conjunto de dados destinado à etapa de teste. Esse processo foi realizado diversas vezes para cada execução, garantindo assim que a cada iteração os dados utilizados na etapa de treinamento fossem diferentes das amostras utilizadas para testar o classificador gerado. Conforme descrito na Seção 4.4, a utilização de conjuntos de dados distintos durante as etapas de treinamento e teste têm como objetivo avaliar a capacidade de generalização (classificação de registros desconhecidos) apresentada pelo algoritmo implementado e identificar possíveis problemas com os dados analisados (tais como *overfitting* e *underfitting*) (WITTEN et al., 2016).

6.5.2 Ajuste de Parâmetros

Grande parte dos algoritmos de ML possuem parâmetros cujos valores devem ser especificados durante a construção do classificador. Esses valores, em geral, influenciam significativamente no desempenho dos algoritmos implementados. Sendo assim, o processo de ajuste desses parâmetros tem como objetivo encontrar os valores mais adequados para as configurações suportadas pelo classificador e, desta forma, obter um melhor desempenho e maior robustez dessas técnicas durante a classificação de um determinado conjunto de dados. A busca pelos melhores valores para os parâmetros geralmente envolve um processo de otimização (que pode ser realizado mediante a utilização de estimadores como, por exemplo, *GridSearchCV*), o que pode contribuir para que o modelo desenvolvido seja custoso computacionalmente. A Tabela 6 apresenta o conjunto de parâmetros utilizado em cada algoritmo.

Tabela 6 – Conjunto de parâmetros utilizados.

Classificador	Parâmetros
Regressão Logística	padrão
K-Nearest Neighbors	algorithm: ‘auto’, leaf_size: ‘30’, n_neighbors: ‘2’
Árvore de Decisão	criterion: ‘gini’ max_depth: ‘None’, max_features: None max_leaf_nodes: 10
Gaussian Naive Bayes	padrão
Máquinas de Vetores Suporte	coef0: 0.0, C: 1, gamma: 0.0001, kernel: ‘rbf’

6.5.3 Avaliação dos Classificadores

Para calcular a eficiência de cada um dos modelos desenvolvidos, foi efetuada uma comparação entre os resultados apresentados por cada algoritmo implementado. A avaliação do desempenho apresentado pelos algoritmos é um procedimento importante para a fase de conclusão desse estudo e foi realizada considerando uma série de indicadores de desempenho (acurácia, precisão, sensibilidade, especificidade, *F-score*, curva ROC e MCC). Esses indicadores foram calculados por meio do uso da matriz de confusão gerada por cada modelo. Sendo assim, para cada classificação de um novo registro realizada pelo modelo, se a saída do classificador fosse igual ao rótulo da amostra, então uma classificação correta foi considerada, caso contrário, um erro foi considerado. A taxa de acerto de cada classificador foi obtida pela razão entre a quantidade de amostras rotuladas corretamente em relação ao número total de registros, ou seja, para cada classificador avaliado uma matriz de confusão foi gerada.

Os modelos desenvolvidos utilizando algoritmos de ML, foram submetidos a um total de 50 execuções. Avaliou-se também o desempenho apresentado por cada algoritmo em termos de tempo de execução. Sendo assim, o tempo de execução é apresentado como a média de 50 execuções. As métricas de desempenho que atestam a capacidade de generalização apresentada por cada modelo serão evidenciadas no Capítulo 7, mediante a utilização de matrizes de confusão, tabelas, gráficos, etc.

7 Resultados e discussões

Neste capítulo são apresentados e discutidos os resultados dos experimentos realizados nesta pesquisa. A Seção 7.1 apresenta as métricas de desempenho obtidas para cada um dos cinco modelos implementados. A Seção 7.2 apresenta uma análise dos resultados obtidos e expõe uma análise comparativa do desempenho apresentado por cada modelo.

7.1 Análise dos resultados

Após a definição dos procedimentos e métodos descritos nos capítulos anteriores, os modelos de análise de confiabilidade de dados de *crowdsourcing* foram construídos utilizando o conjunto de dados mencionado na Seção 6.2. A análise comparativa da capacidade de classificação apresentada por cada um dos modelos desenvolvidos foi realizada considerando os valores das métricas de desempenho descritas na Seção 4.4.1 a saber: matriz de confusão, acurácia, precisão, sensibilidade, especificidade, *F-score*, curva ROC e MCC.

Os resultados analisados e apresentados nas próximas seções descrevem o melhor desempenho obtido pelo modelo durante a fase de teste. Sendo assim, os valores calculados para os gráficos de curvas ROC e AUC, bem como os percentuais derivados da matriz de confusão, são utilizados como métricas de avaliação e comparação do desempenho apresentado por cada modelo implementado.

7.1.1 Modelo com Regressão Logística

A Figura 15 ilustra as métricas de desempenho obtidas durante a execução do modelo desenvolvido utilizando o algoritmo de regressão logística. O classificador gerado por esse algoritmo apresentou uma acurácia total de 97,15%. Analisando a matriz de confusão (Figura 15a) é possível identificar que esse algoritmo foi capaz de prever 99,67% dos registros classificados como confiáveis e 90,94% dos registros considerados como não confiáveis. É possível perceber que o classificador expressa uma boa capacidade de classificar os casos onde os registros são considerados confiáveis. Contudo, esse algoritmo apresenta um desempenho moderadamente inferior durante a classificação dos registros não confiáveis.

Figura 15 – Indicadores de desempenho do algoritmo de regressão logística.

Fonte: Elaborado pelo autor.

Analizando o gráfico de curva ROC (Figura 15b), observa-se que ele apresenta uma AUC de 0,9584, o que significa que o classificador construído apresenta uma excelente capacidade de generalização dos dados. A Figura 15c demonstra ainda, que o modelo obteve indicadores de desempenho com 91% de precisão, 99% de sensibilidade, 96% de especificidade, 95% de *F-score* e 93% de MCC. Conforme destacado na Figura 15e, esse modelo apresentou um tempo de execução médio de 0,047420 segundo.

O modelo desenvolvido utilizando regressão logística, embora apresente bons indicadores, pode ser considerado um modelo mais conservador, pois, apresenta uma moderada predominância em classificar erroneamente os casos de falsos negativos em relação aos casos de falsos positivos. Os resultados das 50 execuções realizadas para esse modelo, se encontram no Apêndice A deste documento.

7.1.2 Modelo com *K-Nearest Neighbour*

O classificador construído por meio do algoritmo *K-Nearest Neighbour*, assim como os outros modelos, apresentou bons resultados, a Figura 16 ilustra os indicadores de desempenho obtidos durante a execução desse modelo. A Figura 16a apresenta a matriz de confusão gerada. A partir da sua análise, é possível verificar que esse modelo conseguiu realizar a identificação correta de 99,67% dos registros classificados como confiáveis e 98,41% dos registros considerados como não confiáveis. Sendo assim, a acurácia total apresentada pelo classificador foi de 99,30%.

Figura 16 – Indicadores de desempenho do modelo de *K-Nearest Neighbour*.

Fonte: Elaborado pelo autor.

A curva ROC, exibida na Figura 16b, indica que este classificador apresenta um excelente poder descritivo sobre o conjunto de dados analisado. O percentual de AUC apresentado por esse modelo foi de 0,9906. Na Figura 16c é possível verificar que o modelo também apresenta bons indicadores de desempenho com 98% de precisão, 99% de sensibilidade, 99% de especificidade, 99% de *F-score* e 98% de MCC. Analisando os

resultados, é possível verificar que, diferentemente do modelo implementado utilizando regressão logística, o classificador gerado por esse modelo não apresenta uma predominância em relação a uma classe específica durante a atividade de classificação dos registros. Além disso, conforme destacado na Figura 16e, o tempo de execução observado para esse modelo também foi relativamente satisfatório, com uma média de 0,024624 segundo. Os resultados obtidos com a execução desse modelo, são descritos no apêndice A.

7.1.3 Modelo com Árvore de Decisão

O modelo desenvolvido utilizando o algoritmo de Árvore de Decisão apresentou desempenho levemente superior ao classificador gerado pelo algoritmo *K-Nearest Neighbour*. A Figura 17 apresenta as métricas de desempenho apresentadas por esse modelo. O classificador obteve uma acurácia total de 99,34%.

Figura 17 – Indicadores de desempenho do modelo de Árvore de Decisão.

Fonte: Elaborado pelo autor.

A matriz de confusão (Figura 17a) demonstra que esse modelo foi capaz de prever 99,55% dos registros classificados como confiáveis e 98,82% dos registros considerados como não confiáveis. Na Figura 17b, a representação gráfica da curva ROC do modelo é exibida. É possível observar que neste modelo o percentual na área sob a curva de 0,9920, o que indica que o classificador construído apresenta uma excelente capacidade de generalização sobre o conjunto de dados em particular.

Analizando os indicadores mostrados na Figura 17c e o tempo de execução descrito pela Figura 17e, observa-se que este modelo foi extremamente eficaz e eficiente em classificar a confiabilidade dos dados de *crowdsourcing*. O classificador gerado obteve 99% de precisão, 99% de sensibilidade, 100% de especificidade, 99% de *F-score* e 98% de MCC com um tempo de execução médio de 0,023074 segundo. Os resultados das execuções realizadas para esse modelo, se encontra no apêndice A.

7.1.4 Modelo com *Gaussian Naive Bayes*

Os indicadores obtidos pela implementação utilizando o algoritmo *Gaussian Naive Bayes* são apresentados na Figura 18. Conforme pode ser observado a partir da matriz de confusão (Figura 18b), esse modelo foi capaz de prever 99,95% dos registros classificados como confiáveis e 93,81% dos registros considerados como não confiáveis. Assim como ocorre no modelo de regressão logística, o classificador apresenta uma boa capacidade de classificar os casos onde os registros são considerados confiáveis, mas apresenta um desempenho moderadamente inferior durante a classificação dos registros não confiáveis.

Figura 18 – Indicadores de desempenho do modelo de *Gaussian Naive Bayes*.

Fonte: Elaborado pelo autor.

O gráfico de curva ROC, mostrado na Figura 18b, demonstra que o modelo construído apresenta uma área sobre a curva de 0,9679. A partir da análise dos dados dos indicadores descritos na Figura 18c e do tempo de execução apresentado pela Figura 18e, é possível observar ainda que o modelo implementado utilizando o algoritmo *Gaussian Naive Bayes* obteve 94% de precisão, 100% de sensibilidade, 98% de especificidade, 97% de *F-score* e 96% de MCC, com um tempo de execução médio de 0,003182 segundo. No apêndice A deste documento, é apresentando os resultados obtidos a partir das 50 execuções realizadas para esse modelo.

7.1.5 Modelo com Máquinas de Vetores Suporte

A Figura 19 ilustra as métricas de desempenho apresentadas pelo modelo implementado utilizando o algoritmo de Máquinas de Vetores Suporte. O classificador gerado por esse modelo apresentou uma acurácia total de 99,09%. Conforme destacado na matriz

de confusão (Figura 19a), o modelo foi capaz de prever corretamente 99,9% dos registros classificados como confiáveis e 97,1% dos registros considerados como não confiáveis.

Figura 19 – Indicadores de desempenho do modelo de Máquinas de Vetores Suporte.

Fonte: Elaborado pelo autor.

A Figura 19b destaca o gráfico de curva ROC do modelo. A área sob a curva calculada para esse modelo obteve um percentual de 0,9832. Conforme apresenta a Figura 19c, os indicadores de desempenho produzidos por esse modelo também apresentaram bons resultados, com 97% de precisão, 100% de sensibilidade, 99% de especificidade, 98% de *F-score* e 98% de MCC. Contudo, conforme destacado na Figura 19e, esse modelo apresentou o maior tempo de execução dentre as técnicas implementadas, com 4,839027 segundos. Os resultados obtidos com a execução desse modelo, são descritos no apêndice A deste documento.

7.2 Comparação dos resultados

O modelo de análise de confiabilidade dos dados obtidos com *crowdsourcing* proposto por essa pesquisa tem como objetivo identificar as contribuições realizadas pelos usuários que, por algum motivo, apresentam informações não confiáveis. Cinco algoritmos de ML foram implementados: (i) Regressão logística, (ii) *K-Nearest Neighbour*, (iii) Árvore de Decisão, (iv) *Gaussian Naive Bayes*, e (v) Máquinas de Vetores Suporte. A Tabela 7 apresenta a média dos resultados de desempenho de 50 execuções de cada um dos algoritmos implementados.

Tabela 7 – Indicadores de desempenho apresentado pelos classificadores.

Classificador	Indicadores de desempenho						
	Precisão	Sensibilidade	Especificidade	F-score	MCC	Acurácia	Tempo médio de execução (s)
Regressão Logística	0,91	0,99	0,96	0,95	0,93	$0,9715 \pm 0,03$	0,047420
K-Nearest Neighbors	0,98	0,99	0,99	0,99	0,98	$0,9930 \pm 0,01$	0,024624
Árvore de Decisão	0,99	0,99	1,00	0,99	0,98	$0,9934 \pm 0,01$	0,023074
Gaussian Naive Bayes	0,94	1,00	0,98	0,97	0,96	$0,9819 \pm 0,02$	0,003182
Máquinas de Vetores Suporte	0,97	1,00	0,99	0,98	0,98	$0,9909 \pm 0,01$	4.839027

Fonte: Elaborado pelo autor.

Analisando as métricas apresentadas por cada modelo, é possível verificar que os classificadores gerados por meio do uso dos algoritmos de ML apresentaram bons resultados. A Figura 20 apresenta uma comparação gráfica entre os valores dos indicadores (exceto o indicador ‘tempo médio de execução’, que não é uma porcentagem) apresentados por cada um dos modelos implementados.

Figura 20 – Comparação dos indicadores de desempenho dos classificadores.

Fonte: Elaborado pelo autor.

O modelo construído utilizando a técnica de regressão logística produziu o menor desempenho em termos de eficácia e eficiência dentre todas as outras técnicas. Esse modelo apresentou a menor capacidade de generalização dos dados, com um percentual de MCC de 93%. O classificador gerado foi efetivo em classificar os casos de registros confiáveis, mas sua habilidade em classificar registros como não confiáveis foi moderadamente inferior ao de outros classificadores. Esse modelo apresentou ainda uma acurácia total de 97,15% e F-score de 95%. O tempo de execução do algoritmo foi bom, apesar de não ser um dos melhores deste estudo, com 0,047420 segundo em média.

Já o modelo implementado com o algoritmo *Gaussian Naive Bayes* apresentou um desempenho um pouco superior ao modelo construído utilizando regressão logística. O modelo desenvolvido apresentou uma boa capacidade de generalização dos dados, com MCC de 96%, acurácia de 98,18% e F-score de 97%. Além disso, dentre as técnicas

implementadas, o classificador gerado por esse modelo foi o que apresentou o melhor tempo de execução, com um tempo médio de 0,003182 segundo.

Os modelos de classificação embasados nos algoritmos de Máquinas de Vetores Suporte, *K-Nearest Neighbour* e Árvore de Decisão obtiveram desempenhos similares. Entretanto, a partir da análise mais aprofundada dos resultados, identificamos que o algoritmo de Árvore de Decisão obteve um leve ganho de desempenho sobre as técnicas de Máquinas de Vetores Suporte e *K-Nearest Neighbour*. O classificador gerado utilizando Árvore de Decisão apresentou excelentes indicadores em todas as métricas de avaliação. O modelo gerado obteve uma acurácia total de 94,34% e F-score de 99%. Esse classificador também apresentou um MCC de 98%, o que indica que o modelo tem um ótima capacidade de discriminar corretamente as classes dos dados. Além disso, esse modelo apresentou o segundo melhor tempo de execução, com um tempo médio de 0,023074 segundo.

Assim como o modelo desenvolvido utilizando o algoritmo de Árvore de Decisão, o modelo baseado em KNN também mostrou-se eficaz e eficiente na realização da classificação de confiabilidade do conjunto de dados obtido mediante a utilização de *crowdsourcing*, com uma acurácia de 99,30%, F-score de 99% e MCC de 98%. O tempo de execução apresentado por esse modelo também pode ser considerado satisfatório, com um tempo médio de 0,024624 segundo. Contudo, a partir da análise dos resultados, é possível identificar que o modelo desenvolvido utilizando o algoritmo de Árvore de Decisão apresenta indicadores de desempenho moderadamente superiores ao modelo implementado utilizando KNN.

Por fim, o classificador implementado utilizando SVM, assim como os outros modelos, também apresentou excelentes. Esse algoritmo apresentou uma acurácia de 99,09%, F-score de 98% e MCC de 96%. Contudo, esse modelo foi o que apresentou o maior tempo de execução dentre todos os modelos implementados neste trabalho, com um tempo médio de execução de 4,839027 segundos. De modo geral, o modelo desenvolvido utilizando o algoritmo de SVM, em termos de eficácia, pode ser considerado similar aos modelos baseados em Árvore de Decisão e KKN, embora, significativamente inferior em termos de eficiência quando comparado com todos os outros modelos implementados.

Diversos trabalhos encontrados na literatura tratam do problema de como utilizar as técnicas de *crowdsourcing* como abordagem complementar de captura de dados para solucionar os problemas relacionados às estruturas de SC (HOAR, 2010; FURTADO et al., 2010; MISRA et al., 2014; MIRRI et al., 2014; ZHOU; ZHENG; LI, 2012; CEROTTI et al., 2016; OLIVEIRA et al., 2016; STERNBERG; LANTZ, 2018). Contudo, nenhum desses

trabalhos investigou o uso de técnicas de aprendizado de máquina para criação de modelos que realizem a análise da confiabilidade desses dados de forma eficiente e automatizada, para, desta forma, mitigar as incertezas introduzidas pelo uso de *crowdsourcing*.

Os resultados experimentais obtidos neste trabalho indicam que técnicas de ML podem ser empregadas para a análise de confiabilidade dos dados sobre a localização dos ônibus da cidade de São Paulo disponibilizados pelos seus passageiros. Os modelos de análise empregados apresentaram bons indicadores de desempenho, alcançando uma boa precisão e eficiência nas atividades de classificação dessas informações.

É importante ressaltar, contudo, que vários fatores podem ter contribuído para que os algoritmos implementados neste trabalho tenham apresentado ótimos resultados. O *dataset* utilizado nesta pesquisa foi submetido a um processo de normalização e, deste modo, o conjunto de dados apresenta poucas inconsistências e propriedades estocásticas que poderiam influenciar significativamente na atividade de classificação dos registros. A construção do modelo foi realizada baseando-se em apenas 4 variáveis descritivas. Além disso, os modelos desenvolvidos foram treinados e testados utilizando um conjunto de dados pequeno, que não reproduz o fluxo de dados obtidos em tempo real. Portanto, é recomendado que em trabalhos futuros esses modelos sejam analisados possivelmente recebendo uma alta carga de informações, para que então seja possível conhecer a real eficiência e eficácia de cada modelo.

8 Conclusão

No contexto de transporte público, um dos problemas existentes é a pontualidade dos ônibus em relação aos horários de parada preestabelecido. A integração eficiente das diversas tecnologias existentes nas SC tem permitido a criação e identificação de inúmeras oportunidades para solucionar (ou amenizar) problemas como o mencionado, que afetam a qualidade dos serviços de transporte disponibilizados para a população.

A utilização de *crowdsourcing* em tese possibilita a obtenção de novos dados ou até mesmo a correção de informações existentes. As informações sobre a localização dos veículos de uma rede de transporte público (modal ônibus), disponibilizadas pelos próprios usuários em tempo real por meio de seus dispositivos móveis, podem ser combinadas às informações coletadas por outros sistemas distribuídos, como, por exemplo, os sistemas AVL instalados nos ônibus, contribuindo, assim, para a criação de SIU mais eficientes e confiáveis. No entanto, os dados de *crowdsourcing* também apresentam informações não confiáveis que podem comprometer a sua confiabilidade.

Não há, de acordo com o levantamento bibliográfico realizado, um modelo que realize de forma automática a análise da confiabilidade de dados de *crowdsourcing* especificamente para o domínio de transporte público. O levantamento bibliográfico indicou, ainda, que embora a confiabilidade dos dados coletados com uso de *crowdsourcing* seja um fator crucial, ainda não existe uma abordagem específica que possibilite avaliar de forma eficiente a qualidade dessas informações.

O objetivo deste trabalho, definido na Seção 1.2, foi propor um modelo para a análise de confiabilidade dos dados obtidos com *crowdsourcing*, que combina um conjunto de métricas escolhidas para o contexto do modal ônibus com técnicas de aprendizado de máquina. Sendo assim, cinco algoritmos de ML foram implementados como parte desta pesquisa.

A hipótese aqui avaliada é a de que o modelo proposto pode assegurar a confiabilidade dos dados obtidos voluntariamente com o auxílio de *crowdsourcing* sobre a localização de ônibus, com um bom desempenho para ser implementado em um SIU.

A avaliação dessa hipótese consistiu em 50 execuções para cada um dos cinco algoritmos de ML implementados. Cada algoritmo foi avaliado utilizando um conjunto de dados final contendo 13,493 (treze mil e quatrocentos e noventa e três) registros. Sendo

assim, faz-se pertinente uma análise comparativa, de modo a encontrar um algoritmo mais eficiente para ser aplicado em tempo real e eficaz para analisar a confiabilidade de dados obtidos via *crowdsourcing*.

Os resultados expostos no capítulo anterior indicam que os modelos desenvolvidos apresentam bons resultados. A análise dos resultados mostra ainda, que o modelo desenvolvido utilizando o algoritmo baseado em Árvore de Decisão foi extremamente eficaz e eficiente na realização da atividade de classificação dos dados obtidos com *crowdsourcing*. Dentre os modelos analisados, foi o que apresentou os melhores indicadores de desempenhos em termos de acurácia (94,34%) e F-score (99%), e o segundo melhor tempo de execução (0,023074 segundo). Já modelo implementado utilizando o Gaussian Naive Bayes também obteve um bom desempenho. Dentre os modelos desenvolvidos, foi o mais eficiente, com um tempo de execução médio de 0,003182 segundo. Contudo, embora tenha apresentado bons indicadores em termos de eficácia, esse modelo apresentou o quarto melhor resultado em termos acurácia (98,18%) e F-score (97%).

Acreditamos que a qualidade da classificação e o bom desempenho computacional apresentado por esses algoritmos os tornam bons candidatos a serem utilizados por um sistema de coleta de dados em tempo real. Note que o algoritmo de Árvore de Decisão destacou-se pelo seu desempenho em termos de eficácia e eficiência, enquanto o algoritmo de Gaussian Naive Bayes destacou-se pela eficiência com o melhor tempo de execução.

8.1 Contribuições

A contribuição deste trabalho de mestrado é o desenvolvimento de um método para analisar a confiabilidade dos dados obtidos com o auxílio de *crowdsourcing*, aplicado à localização dos ônibus da rede de transporte público. Para o desenvolvimento desse método foi proposta uma abordagem específica para o tratamento e análise dos dados coletados especializados para o sistema de transporte público (por ônibus) do município de São Paulo. Foi proposto também um protocolo com heurísticas específicas para a atividade de classificação da confiabilidade desses dados.

Durante a realização desse estudo foram realizados vários experimentos aplicando-se os métodos propostos ao problema de análise da confiabilidade das informações de *crowdsourcing*. Acreditamos que a futura incorporação desses modelos em um sistema de

informação ao usuário será também uma contribuição à sociedade. O aumento da qualidade dessas informações possui potencial para melhorar a qualidade dos serviços de transporte disponibilizados para a população.

8.2 *Trabalho publicado*

As contribuições científicas originadas diretamente desta pesquisa resultaram em uma publicação:

- Diego Vieira Neves, Felipe Cordeiro Alves Dias e Daniel Cordeiro. Uso de aprendizado supervisionado para análise de confiabilidade de dados de *crowdsourcing* sobre posicionamento de ônibus. I Workshop Brasileiro de Cidades Inteligentes (WBCI 2018), Natal-RN, Brasil. Anais do XXXVIII Congresso da Sociedade Brasileira de Computação (CSBC 2018).

8.3 *Trabalhos futuros*

O projeto de pesquisa aqui apresentado pode ser estendido de diversas formas. Pretende-se, portanto, desenvolver as seguintes atividades como trabalhos futuros:

- Incorporar os modelos de análise de confiabilidade de dados de *crowdsourcing* desenvolvidos neste trabalho em um sistema de informação ao usuário;
- Analisar o desempenho computacional e a qualidade de classificação apresentada pelos algoritmos implementados, na análise de fluxos de dados obtidos em tempo real para comprovar a sua eficiência e eficácia;
- Incorporar ao conjunto de dados utilizados, informações referentes à localização dos veículos de uma rede de transporte público de outros municípios;
- Analisar a confiabilidade de outras formas de obtenção de dados como, por exemplo, informações provenientes de diversos sensores e mídias sociais.

Referências¹

- ALLAHBAKHS, M. et al. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 17, n. 2, p. 76–81, mar. 2013. ISSN 1089-7801. Disponível em: <http://dx.doi.org/10.1109/MIC.2013.20>. Citado na página 47.
- ALPAYDIN, E. *Introduction to machine learning*. 3^a. ed. Cambridge, Massachusetts: MIT press, 2014. Citado na página 29.
- ANTP. *Sistema de Informações da Mobilidade Urbana: Relatório Geral 2014*. 2016. Citado na página 20.
- BARBEAU, S. J.; BORNING, A.; WATKINS, K. Onebusaway multi-region—rapidly expanding mobile transit apps to new cities. *Journal of Public Transportation*, v. 17, n. 4, p. 3, 2014. Citado na página 27.
- BATAGAN, L. Smart cities and sustainability models. *Informatica Economica*, Academy of Economic Studies-Bucharest, Romania, v. 15, n. 3, p. 80–87, 2011. Citado 2 vezes nas páginas 18 e 19.
- BATISTA, D. M. et al. Interscity: Addressing future internet research challenges for smart cities. 7th IEEE International Conference on Network of the Future, 2016. Citado 2 vezes nas páginas 14 e 18.
- BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. Citado 2 vezes nas páginas 30 e 31.
- BRABHAM, D. C. *Crowdsourcing*. Cambridge, Massachusetts: Mit Press, 2013. Citado 3 vezes nas páginas 15, 24 e 25.
- BREIMAN, L. et al. *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks Cole Advanced Books and Software, 1984. Citado na página 35.
- BRINK, H.; RICHARDS, J. W.; FETHEROLF, M. *Real-world machine learning*. 2^a. ed. Shelter Island, New York: Manning Publications, 2017. Citado na página 41.
- CAPROTTI, F. *Eco-cities and the transition to low carbon economies*. London, England: Palgrave Pivot, 2016. Citado na página 18.
- CEROTTI, D. et al. A crowd-cooperative approach for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, IEEE, 2016. Citado 10 vezes nas páginas 15, 20, 23, 24, 25, 43, 44, 47, 48 e 71.
- CHOWDHURY, M. A.; SADEK, A. W. *Fundamentals of intelligent transportation systems planning*. London, England: Artech House, 2003. Citado na página 21.
- CULLINA, E.; CONBOY, K.; MORGAN, L. Measuring the crowd: a preliminary taxonomy of crowdsourcing metrics. In: ACM. *Proceedings of the 11th International Symposium on Open Collaboration*. New York, NY, USA, 2015. p. 7. Citado 2 vezes nas páginas 24 e 25.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

CUNHA, J. N. *Metodologia de geração dinâmica de padrões de viagens rodoviárias para monitoramentos inteligentes de veículos de carga em sistemas AVL*. Tese (Doutorado) — Universidade de São Paulo, 2008. Citado na página 22.

CURY, M. J. F.; MARQUES, J. A. L. F. A cidade inteligente: uma reterritorialização smart city: A reterritorialization. *Redes*, v. 22, n. 1, 2016. Citado na página 18.

DANIEL, F. et al. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 51, n. 1, p. 7:1–7:40, jan. 2018. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/3148148>. Citado na página 47.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification (2Nd Edition)*. New York, NY, USA: Wiley-Interscience, 2000. ISBN 0471056693. Citado na página 35.

FAWCETT, T. An introduction to ROC analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado 2 vezes nas páginas 39 e 41.

FERRAZ, A. C. P.; TORRES, I. G. E. *Transporte Público Urbano*. 2ª. ed. São Paulo - SP: ED-RIMA, 2004. Citado 2 vezes nas páginas 19 e 42.

FERRI, C.; HERNÁNDEZ-ORALLO, J.; MODROIU, R. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, Elsevier, v. 30, n. 1, p. 27–38, 2009. Citado 2 vezes nas páginas 37 e 39.

FERRIS, B.; WATKINS, K.; BORNING, A. Onebusaway: A transit traveler information system. In: SPRINGER. *International Conference on Mobile Computing, Applications, and Services*. Berlin, Heidelberg, 2009. p. 92–106. Citado na página 27.

FERRIS, B.; WATKINS, K.; BORNING, A. Location-aware tools for improving public transit usability. *IEEE Pervasive Computing*, v. 9, n. 1, p. 13–19, 2010. Citado na página 19.

FERRIS, B.; WATKINS, K.; BORNING, A. Onebusaway: Behavioral and satisfaction changes resulting from providing real-time arrival information for public transit. *Proceedings of the 2011 Transportation research board annual meeting*, 2011. Citado na página 27.

FIGUEIREDO, G. D. S. *Técnicas de Simulação para Apoio à Decisão em Planejamento Urbano*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2016. Citado 5 vezes nas páginas 14, 15, 18, 21 e 26.

FIGUEIREDO, L. M. B. *Sistemas inteligentes de transporte*. Tese (Doutorado) — Universidade do Porto, 2005. Citado na página 21.

FURTADO, V. et al. Collective intelligence in law enforcement—the wikicrimes system. *Information Sciences*, Elsevier, v. 180, n. 1, p. 4–17, 2010. Citado 3 vezes nas páginas 43, 46 e 71.

GHOTRA, B.; MCINTOSH, S.; HASSAN, A. E. Revisiting the impact of classification techniques on the performance of defect prediction models. In: IEEE PRESS. *Proceedings of the 37th International Conference on Software Engineering-Volume 1*. Florence, Italy, 2015. p. 789–800. Citado 6 vezes nas páginas 29, 30, 31, 32, 37 e 41.

GOMEZ, A.; DIAZ, G.; BOUSSETTA, K. Its forecast: Gis integration with active sensory system. In: IEEE. *Information Infrastructure Symposium*. GIIS'09, 2009. p. 1–6. Citado na página 21.

HAND, D. J. *Construction and assessment of classification rules*. New York: Wiley Chichester, 1997. v. 15. Citado na página 41.

HANDWERK, B. *Half of humanity will live in cities by year's end*. 2008. Disponível em: <https://news.nationalgeographic.com/news/2008/03/080313-cities.html>. Citado na página 20.

HOAR, R. A personalized web based public transit information system with user feedback. In: IEEE. *13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. Funchal, Portugal, 2010. p. 1807–1812. Citado 5 vezes nas páginas 43, 46, 47, 48 e 71.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. New York, NY, USA: John Wiley & Sons, 2013. v. 398. Citado 3 vezes nas páginas 33, 41 e 42.

HOWE, J. The rise of crowdsourcing. *Wired magazine*, v. 14, n. 6, p. 1–4, 2006. Citado na página 24.

HSU, C. wei; CHANG, C. chung; LIN, C. jen. A practical guide to support vector classification. Taipei, 2010. Citado na página 36.

JAGABATHULA, S.; SUBRAMANIAN, L.; VENKATARAMAN, A. Reputation-based worker filtering in crowdsourcing. In: *Advances in Neural Information Processing Systems*. New York, NY, USA: Curran Associates, Inc., 2014. p. 2492–2500. Citado na página 47.

JAMES, G. et al. *An introduction to statistical learning*. New York, NY, USA: Springer, 2013. v. 112. Disponível em: <https://doi.org/10.1007/978-1-4614-7138-7>. Citado 2 vezes nas páginas 37 e 38.

JEONG, R.; RILETT, R. Bus arrival time prediction using artificial neural network model. In: IEEE. *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems*. [S.l.], 2004. p. 988–993. Citado 2 vezes nas páginas 14 e 22.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. v. 14, n. 2, p. 1137–1145, 1995. Citado 3 vezes nas páginas 38, 39 e 60.

MACEDO, E. L. C. *Previsão de Tráfego em Enlaces de Redes Utilizando Séries Temporais*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2015. Citado 2 vezes nas páginas 54 e 56.

MAGALHÃES, C. T. d. A. *Avaliação de Tecnologias de Rastreamento por GPS para o Monitoramento do Transporte Público por Ônibus*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, 2008. Citado 2 vezes nas páginas 14 e 21.

MARSLAND, S. *Machine learning: an algorithmic perspective*. 2. ed. New York: CRC press, 2015. Citado 4 vezes nas páginas 29, 30, 31 e 32.

MASHHADI, A. J.; CAPRA, L. Quality control for real-time ubiquitous crowdsourcing. In: *Proceedings of the 2nd International Workshop on Ubiquitous Crowdsourcing*. New York, NY, USA: ACM, 2011. (UbiCrowd '11), p. 5–8. ISBN 978-1-4503-0927-1. Disponível em: <http://doi.acm.org/10.1145/2030100.2030103>. Citado na página 47.

MIRRI, S. et al. On combining crowdsourcing, sensing and open data for an accessible smart city. In: IEEE. *Next Generation Mobile Apps, Services and Technologies (NGMAST), 2014 Eighth International Conference on*. Oxford, UK, 2014. p. 294–299. Citado 3 vezes nas páginas 43, 45 e 71.

MISRA, A. et al. Crowdsourcing and its application to transportation data collection and management. *Transportation Research Record: Journal of the Transportation Research Board*, Transportation Research Board of the National Academies, n. 2414, p. 1–8, 2014. Citado 12 vezes nas páginas 15, 20, 23, 24, 25, 26, 43, 44, 45, 47, 48 e 71.

MITCHELL, T. M. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, v. 45, n. 37, p. 870–877, 1997. Citado 9 vezes nas páginas 29, 30, 31, 32, 33, 34, 36, 37 e 38.

MOUSA, H. et al. Trust management and reputation systems in mobile participatory sensing applications. *Comput. Netw.*, Elsevier North-Holland, Inc., New York, NY, USA, v. 90, n. C, p. 49–73, out. 2015. ISSN 1389-1286. Disponível em: <https://doi.org/10.1016/j.comnet.2015.07.011>. Citado na página 47.

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*. 1. ed. Cambridge, Massachusetts: MIT press, 2012. Citado na página 29.

OLIVEIRA, R. et al. Gamificação e crowdsourcing no combate sustentável ao aedes aegypti. In: SBIE. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação)*. Uberlândia-MG, Brasil, 2016. v. 27, n. 1, p. 390. Citado 3 vezes nas páginas 43, 44 e 71.

PAGANI, A. et al. Reconstruction of public transport state. In: IEEE. *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. Rio de Janeiro-RJ, Brasil, 2016. p. 2285–2292. Citado 2 vezes nas páginas 26 e 27.

PAPAGIANNAKI, K. et al. Long-term forecasting of internet backbone traffic: Observations and initial models. In: IEEE. *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*. San Francisco, CA, USA, 2003. v. 2, p. 1178–1188. Citado 2 vezes nas páginas 54 e 56.

PEDERSEN, J. et al. Conceptual foundations of crowdsourcing: A review of its research. In: IEEE. *System Sciences (HICSS), 2013 46th Hawaii International Conference on*. Wailea, Maui, HI, USA, 2013. p. 579–588. Citado 3 vezes nas páginas 15, 24 e 25.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 58.

PILON, J. Sistema de informação ao usuário do transporte coletivo por ônibus na cidade de vitória-ES. *Ponta Grossa: Universidade Tecnológica Federal do Paraná*, 2009. Citado na página 22.

- PONS, I.; MONTEIRO, J.; CARDOSO, R. S. *Big Data para análise de métricas de qualidade de transporte: metodologia e aplicação*. São Paulo-SP, Brasil: ANTP, 2015. Citado 6 vezes nas páginas 14, 20, 21, 22, 23 e 55.
- QUINLAN, J. R. C4.5: Programming for machine learning. *Morgan Kaufmann Publishers Inc.*, San Francisco, CA, USA, v. 38, p. 48, 1993. Citado na página 35.
- SAMSEL, C. et al. Cascading information for public transport assistance. In: WEBIST 2014. *Proceedings of the 10th International Conference on Web Information Systems and Technologies*. Barcelona, Spain, 2014. v. 1. Citado na página 27.
- SANTIS, R. D. et al. Smart city: fact and fiction. Munich Personal RePEc Archive(MPRA), 2014. Citado na página 14.
- SCHOLKOPF, B.; SMOLA, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. London, England: MIT press, 2001. Citado na página 36.
- SCHWARTZ, C. et al. Modeling crowdsourcing platforms to enable workforce dimensioning. In: IEEE. *Telecommunication Networks and Applications Conference (ITNAC), 2015 International*. Sydney, NSW, Australia, 2015. p. 30–37. Citado 3 vezes nas páginas 24, 25 e 26.
- SHAPIRO, J. M. Smart cities: quality of life, productivity, and the growth effects of human capital. *The review of economics and statistics*, MIT Press, v. 88, n. 2, p. 324–335, 2006. Citado na página 18.
- SILVA, D. M. d. *Sistemas Inteligentes no Transporte Público por Ônibus. 129f*. Tese (Doutorado) — Dissertação (Mestrado em Engenharia de Produção)—Escola de Engenharia, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2000. Citado na página 23.
- SPTRANS. *Sistemas Informatizados para a Gestão do Transporte Coletivo do Município de São Paulo*. 2009. Disponível em: http://www.sptrans.com.br/pdf/biblioteca_tecnica/sistemas_informatizados_para_a_gestao_do_transporte.pdf. Citado na página 23.
- STEENBRUGGEN, J.; TRANOS, E.; NIJKAMP, P. Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, Elsevier, v. 39, n. 3, p. 335–346, 2015. Citado 2 vezes nas páginas 14 e 23.
- STERNBERG, H.; LANTZ, B. Using crowdsourced data to analyze transport crime. *International Journal of Logistics Research and Applications*, Taylor & Francis, v. 21, n. 2, p. 133–147, 2018. Citado 3 vezes nas páginas 43, 44 e 71.
- SUSSMAN, J. S. *Perspectives on intelligent transportation systems (ITS)*. [S.l.]: Springer Science & Business Media, 2008. Citado 2 vezes nas páginas 20 e 21.
- TOPPETA, D. The smart city vision: how innovation and ict can build smart, “livable”, sustainable cities. *The Innovation Knowledge Foundation. Think*, 2010. Citado 2 vezes nas páginas 18 e 19.
- TORREGROSA, G. M.; IGARTUA, M. A.; PUGLISI, S. Improvement of algorithms to identify transportation modes for mobilitapp, an android application to anonymously track citizens in barcelona. *arXiv preprint arXiv:1605.05342*, 2016. Citado na página 26.

TOWNSEND, A. M. *Smart cities: Big data, civic hackers, and the quest for a new utopia*. 1^a. ed. New York: W. W. Norton & Company, 2013. Citado na página 14.

VAPNIK, V. N. *Statistical learning theory*. New York, NY, USA: Wiley-Interscience, 1998. v. 1. Citado na página 36.

WANG, J.; SHI, Q. Short-term traffic speed forecasting hybrid model based on chaos-wavelet analysis-support vector machine theory. *Transportation Research Part C: Emerging Technologies*, Elsevier, v. 27, p. 219–232, 2013. Citado 2 vezes nas páginas 54 e 56.

WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. New York, NY, USA: Morgan Kaufmann, 2016. Citado 6 vezes nas páginas 34, 36, 37, 38, 39 e 60.

WOLFRAM, M. Deconstructing smart cities: an intertextual reading of concepts and practices for integrated urban and ict development. In: REAL CORP. *17th International Conference on Urban Planning, Regional Development and Information Society*. Austria, 2012. Citado na página 20.

ZHOU, P.; ZHENG, Y.; LI, M. How long to wait?: Predicting bus arrival time with mobile phone based participatory sensing. In: *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*. New York, NY, USA: ACM, 2012. (MobiSys '12), p. 379–392. Disponível em: <http://doi.acm.org/10.1145/2307636.2307671>. Citado 5 vezes nas páginas 43, 45, 47, 48 e 71.

ZHU, T. et al. The prediction of bus arrival time using global positioning system data and dynamic traffic information. In: IEEE. *Wireless and Mobile Networking Conference (WMNC), 2011 4th Joint IFIP*. [S.l.], 2011. p. 1–5. Citado 2 vezes nas páginas 14 e 23.

ZHU, W. et al. Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, Baltimore, Maryland, USA, v. 19, p. 67, 2010. Citado na página 41.

Apêndices

Apêndice A – Exemplo de apêndice

Esta seção descreve com detalhes as métricas de desempenho apresentadas pelos cinco algoritmos (Regressão logística, *K-Nearest Neighbour*, Árvore de Decisão, *Gaussian Naive Bayes*, e (*v*) Máquinas de Vetores Suporte) implementados. As Tabelas 8, 9, 10, 11 e 12 descrevem os resultados das 50 execuções realizadas para cada desses algoritmos.

Tabela 8 – Execuções do modelo utilizando regressão logística.

Execução	Matriz de confusão				Métricas de desempenho					Desempenho do treinamento		Desempenho do teste	
	TP	FP	TN	FN	Precisão	Sensibilidade	Especificidade	<i>F-score</i>	MCC	Acurácia	Tempo de execução (s)	Acurácia	Tempo médio de execução (s)
0	1079	109	2850	10	0,91	0,99	0,96	0,95	0,93	0,9727	0,617157	0,9706	0,033884
1	976	107	2956	9	0,90	0,99	0,97	0,94	0,93	0,9701	0,382016	0,9713	0,042845
2	1056	111	2875	6	0,90	0,99	0,96	0,95	0,93	0,9720	0,414698	0,9711	0,037426
3	1055	97	2887	9	0,92	0,99	0,97	0,95	0,94	0,9716	0,327952	0,9738	0,037687
4	1055	113	2873	7	0,90	0,99	0,96	0,95	0,93	0,9731	0,411475	0,9704	0,042967
5	1069	113	2855	11	0,90	0,99	0,96	0,95	0,93	0,9723	0,359711	0,9694	0,141373
6	1106	98	2834	10	0,92	0,99	0,97	0,95	0,94	0,9699	0,895103	0,9733	0,037180
7	1077	118	2845	8	0,90	0,99	0,96	0,94	0,93	0,9724	0,329065	0,9689	0,035372
8	1010	114	2916	8	0,90	0,99	0,96	0,94	0,92	0,9740	0,299749	0,9699	0,030596
9	1035	111	2897	5	0,90	1,00	0,96	0,95	0,93	0,9018	0,450272	0,9713	0,032555
10	1068	109	2863	8	0,91	0,99	0,96	0,95	0,93	0,9731	0,323992	0,9711	0,034014
11	1065	108	2869	6	0,91	0,99	0,96	0,95	0,93	0,9715	0,327651	0,9718	0,034621
12	1078	101	2862	7	0,91	0,99	0,97	0,95	0,94	0,9709	0,340595	0,9733	0,032337
13	1086	90	2859	13	0,92	0,99	0,97	0,95	0,94	0,9711	0,307765	0,9746	0,032713
14	1098	78	2861	11	0,93	0,99	0,97	0,96	0,95	0,9683	0,351396	0,9780	0,033137
15	1054	120	2869	5	0,90	1,00	0,96	0,94	0,93	0,9714	0,316675	0,9691	0,035054
16	1100	107	2828	13	0,91	0,99	0,96	0,95	0,93	0,9723	0,308566	0,9704	0,049525
17	1046	109	2881	12	0,91	0,99	0,96	0,95	0,93	0,9717	0,323621	0,9701	0,049142
18	1077	120	2849	2	0,90	1,00	0,96	0,95	0,93	0,9712	0,331250	0,9699	0,039972
19	1046	101	2891	10	0,91	0,99	0,97	0,95	0,93	0,9700	0,309239	0,9726	0,032338
20	1060	110	2864	14	0,91	0,99	0,96	0,94	0,93	0,9734	0,380652	0,9694	0,032968
21	1068	89	2880	11	0,92	0,99	0,97	0,96	0,94	0,9691	0,328724	0,9753	0,033813
22	1058	115	2869	6	0,90	0,99	0,96	0,95	0,93	0,9715	0,325871	0,9701	0,034468
23	1085	95	2857	11	0,92	0,99	0,97	0,95	0,94	0,9719	0,367051	0,9738	0,033242
24	1080	115	2846	7	0,90	0,99	0,96	0,95	0,93	0,9728	0,341593	0,9699	0,032006
25	1040	116	2882	10	0,90	0,99	0,96	0,94	0,92	0,9726	0,337271	0,9689	0,047695
26	1034	105	2900	9	0,91	0,99	0,97	0,95	0,93	0,9717	0,371388	0,9718	0,038926
27	1050	119	2875	4	0,90	1,00	0,96	0,94	0,93	0,9706	0,343725	0,9696	0,032850
28	1055	118	2872	3	0,90	1,00	0,96	0,95	0,93	0,9715	0,362018	0,9701	0,037926
29	1097	83	2863	5	0,93	1,00	0,97	0,96	0,95	0,9693	0,341616	0,9783	0,035179
30	1070	104	2860	14	0,91	0,99	0,96	0,95	0,93	0,9709	0,491163	0,9708	0,067961
31	1074	105	2858	11	0,91	0,99	0,96	0,95	0,93	0,9712	0,528152	0,9713	0,033979
32	1103	106	2833	6	0,91	0,99	0,96	0,95	0,93	0,9714	0,362948	0,9723	0,037050
33	1067	100	2874	7	0,91	0,99	0,97	0,95	0,94	0,9705	0,325672	0,9736	0,033458
34	1078	120	2844	6	0,90	0,99	0,96	0,94	0,93	0,9725	0,358192	0,9689	0,032215
35	1074	91	2876	7	0,92	0,99	0,97	0,96	0,94	0,9706	0,352186	0,9758	0,037856
36	1054	111	2871	12	0,90	0,99	0,96	0,94	0,93	0,9726	0,334167	0,9696	0,034725
37	1026	106	2908	8	0,91	0,99	0,96	0,95	0,93	0,9714	0,360424	0,9718	0,033766
38	1047	112	2883	6	0,90	0,99	0,96	0,95	0,93	0,9719	0,320201	0,9708	0,034436
39	1069	99	2869	11	0,92	0,99	0,97	0,95	0,93	0,9719	0,325532	0,9728	0,043015
40	1079	110	2847	12	0,91	0,99	0,96	0,95	0,93	0,9731	0,314481	0,9699	0,032465
41	1072	101	2866	9	0,91	0,99	0,97	0,95	0,93	0,9713	0,348112	0,9728	0,032340
42	1072	117	2844	15	0,90	0,99	0,96	0,94	0,92	0,9727	0,362871	0,9674	0,032181
43	1070	109	2865	4	0,91	1,00	0,96	0,95	0,93	0,9716	0,317248	0,9721	0,032595
44	1109	88	2834	17	0,93	0,98	0,97	0,95	0,94	0,9704	0,370378	0,9741	0,053475
45	1043	117	2882	6	0,90	0,99	0,96	0,94	0,93	0,9711	0,498850	0,9696	0,101592
46	1059	106	2877	6	0,91	0,99	0,96	0,95	0,93	0,9725	0,701624	0,9723	0,037566
47	1036	106	2894	12	0,91	0,99	0,96	0,95	0,93	0,9724	0,376070	0,9708	0,035191
48	1065	117	2861	5	0,90	1,00	0,96	0,95	0,93	0,9713	0,354905	0,9699	0,039140
49	1067	93	2879	9	0,92	0,99	0,97	0,95	0,94	0,9694	0,362982	0,9748	0,034925

Tabela 9 – Execuções do modelo utilizando *K-Nearest Neighbour*.

Execução	Matriz de confusão				Métricas de desempenho					Desempenho do treinamento		Desempenho do teste	
	TP	FP	TN	FN	Precisão	Sensibilidade	Especificidade	<i>F-score</i>	MCC	Acurácia	Tempo de execução (s)	Acurácia	Tempo médio de execução (s)
0	1147	18	2870	13	0,98	0,99	0,99	0,99	0,98	0,9927	0,194204	0,9923	0,046634
1	1180	19	2839	10	0,98	0,99	0,99	0,99	0,98	0,9926	0,091898	0,9928	0,020378
2	1131	16	2895	6	0,99	0,99	0,99	0,99	0,99	0,9920	0,086817	0,9946	0,020278
3	1143	22	2874	9	0,98	0,99	0,99	0,99	0,98	0,9921	0,091012	0,9923	0,021480
4	1140	21	2877	10	0,98	0,99	0,99	0,99	0,98	0,9942	0,086993	0,9923	0,019176
5	1200	14	2824	10	0,99	0,99	1,00	0,99	0,99	0,9917	0,092492	0,9941	0,021112
6	1194	24	2823	7	0,98	0,99	0,99	0,99	0,98	0,9935	0,089093	0,9923	0,019949
7	1162	24	2851	11	0,98	0,99	0,99	0,99	0,98	0,9931	0,101766	0,9914	0,019738
8	1166	26	2849	7	0,98	0,99	0,99	0,99	0,98	0,9934	0,092864	0,9918	0,019467
9	1140	22	2879	7	0,98	0,99	0,99	0,99	0,98	0,9923	0,089765	0,9928	0,029539
10	1161	22	2854	11	0,98	0,99	0,99	0,99	0,98	0,9932	0,096738	0,9918	0,020700
11	1136	20	2884	8	0,98	0,99	0,99	0,99	0,98	0,9926	0,098967	0,9931	0,019843
12	1142	11	2883	12	0,99	0,99	1,00	0,99	0,99	0,9922	0,111693	0,9943	0,024709
13	1105	21	2917	5	0,98	1,00	0,99	0,99	0,98	0,9920	0,167090	0,9936	0,035614
14	1138	18	2884	8	0,98	0,99	0,99	0,99	0,98	0,9930	0,163106	0,9936	0,024208
15	1139	15	2887	7	0,99	0,99	0,99	0,99	0,99	0,9929	0,135363	0,9946	0,032304
16	1125	20	2896	7	0,98	0,99	0,99	0,99	0,98	0,9925	0,210401	0,9933	0,025136
17	1125	14	2898	11	0,99	0,99	1,00	0,99	0,98	0,9927	0,118308	0,9938	0,029255
18	1147	21	2868	12	0,98	0,99	0,99	0,99	0,98	0,9926	0,114655	0,9918	0,023683
19	1140	21	2875	12	0,98	0,99	0,99	0,99	0,98	0,9940	0,096591	0,9918	0,024318
20	1167	19	2848	14	0,98	0,99	0,99	0,99	0,98	0,9928	0,098096	0,9918	0,023085
21	1136	17	2887	8	0,99	0,99	0,99	0,99	0,98	0,9935	0,097696	0,9938	0,020004
22	1139	27	2877	5	0,98	1,00	0,99	0,99	0,98	0,9938	0,084203	0,9921	0,020646
23	1121	20	2897	10	0,98	0,99	0,99	0,99	0,98	0,9930	0,085749	0,9926	0,019742
24	1131	17	2889	11	0,99	0,99	0,99	0,99	0,98	0,9915	0,088800	0,9931	0,020364
25	1169	22	2844	13	0,98	0,99	0,99	0,99	0,98	0,9931	0,089680	0,9914	0,022488
26	1134	27	2882	5	0,98	1,00	0,99	0,99	0,98	0,9923	0,094445	0,9921	0,020217
27	1136	10	2891	11	0,99	0,99	1,00	0,99	0,99	0,9920	0,083456	0,9948	0,020416
28	1180	22	2835	11	0,98	0,99	0,99	0,99	0,98	0,9928	0,081374	0,9918	0,021203
29	1143	18	2880	7	0,98	0,99	0,99	0,99	0,98	0,9928	0,104640	0,9938	0,023460
30	1153	21	2865	9	0,98	0,99	0,99	0,99	0,98	0,9922	0,141652	0,9926	0,027316
31	1113	19	2905	11	0,98	0,99	0,99	0,99	0,98	0,9926	0,176610	0,9926	0,020746
32	1132	24	2882	10	0,98	0,99	0,99	0,99	0,98	0,9930	0,085246	0,9916	0,019643
33	1211	14	2817	6	0,99	1,00	1,00	0,99	0,99	0,9922	0,089524	0,9951	0,021190
34	1129	22	2887	10	0,98	0,99	0,99	0,99	0,98	0,9929	0,088083	0,9921	0,019736
35	1170	13	2857	8	0,99	0,99	1,00	0,99	0,99	0,9924	0,082929	0,9948	0,019408
36	1130	18	2888	12	0,98	0,99	0,99	0,99	0,98	0,9926	0,085665	0,9926	0,019311
37	1167	20	2854	7	0,98	0,99	0,99	0,99	0,98	0,9927	0,086736	0,9933	0,020407
38	1135	24	2883	6	0,98	0,99	0,99	0,99	0,98	0,9927	0,075043	0,9926	0,016171
39	1163	19	2852	14	0,98	0,99	0,99	0,99	0,98	0,9922	0,091603	0,9918	0,019863
40	1142	21	2880	5	0,98	1,00	0,99	0,99	0,98	0,9926	0,084533	0,9936	0,018770
41	1125	20	2893	10	0,98	0,99	0,99	0,99	0,98	0,9930	0,095014	0,9926	0,019643
42	1126	18	2893	11	0,98	0,99	0,99	0,99	0,98	0,9916	0,108757	0,9928	0,102179
43	1144	25	2871	8	0,98	0,99	0,99	0,99	0,98	0,9932	0,121809	0,9918	0,028240
44	1148	21	2867	12	0,98	0,99	0,99	0,99	0,98	0,9926	0,085577	0,9918	0,019838
45	1129	23	2891	5	0,98	1,00	0,99	0,99	0,98	0,9926	0,084033	0,9931	0,019713
46	1149	20	2873	6	0,98	0,99	0,99	0,99	0,98	0,9928	0,082069	0,9936	0,018818
47	1138	17	2885	8	0,99	0,99	0,99	0,99	0,98	0,9917	0,085282	0,9938	0,019286
48	1170	20	2855	3	0,98	1,00	0,99	0,99	0,99	0,9938	0,092415	0,9943	0,020737
49	1136	14	2879	19	0,99	0,98	1,00	0,99	0,98	0,9926	0,082950	0,9918	0,020594

Tabela 10 – Execuções do modelo utilizando Árvore de Decisão.

Matriz de confusão					Métricas de desempenho				Desempenho do treinamento			Desempenho do teste	
Execução	TP	FP	TN	FN	Precisão	Sensibilidade	Especificidade	<i>F-score</i>	MCC	Acurácia	Tempo de execução (s)	Acurácia	Tempo médio de execução (s)
0	1138	15	2876	19	0,99	0,98	0,99	0,99	0,98	0,9928	0,270557	0,9916	0,017616
1	1145	15	2878	10	0,99	0,99	0,99	0,99	0,98	0,9924	0,208843	0,9938	0,022232
2	1180	6	2850	12	0,99	0,99	1,00	0,99	0,99	0,9918	0,172546	0,9956	0,014485
3	1151	16	2865	16	0,99	0,99	0,99	0,99	0,98	0,9928	0,150322	0,9921	0,015302
4	1122	18	2893	15	0,98	0,99	0,99	0,99	0,98	0,9942	0,169368	0,9918	0,019844
5	1118	25	2890	15	0,98	0,99	0,99	0,98	0,98	0,9936	0,174900	0,9901	0,019097
6	1161	18	2853	16	0,98	0,99	0,99	0,99	0,98	0,9925	0,190109	0,9916	0,020869
7	1141	21	2870	16	0,98	0,99	0,99	0,98	0,98	0,9930	0,264102	0,9909	0,015763
8	1141	19	2878	10	0,98	0,99	0,99	0,99	0,98	0,9926	0,143803	0,9928	0,014666
9	1159	15	2858	16	0,99	0,99	0,99	0,99	0,98	0,9940	0,175759	0,9923	0,020053
10	1154	19	2863	12	0,98	0,99	0,99	0,99	0,98	0,9921	0,176005	0,9923	0,021224
11	1159	14	2861	14	0,99	0,99	1,00	0,99	0,98	0,9928	0,189247	0,9931	0,017865
12	1144	17	2872	15	0,99	0,99	0,99	0,99	0,98	0,9934	0,164992	0,9921	0,015718
13	1150	15	2871	12	0,99	0,99	0,99	0,99	0,98	0,9930	0,169500	0,9933	0,018264
14	1165	11	2856	16	0,99	0,99	1,00	0,99	0,98	0,9922	0,181179	0,9933	0,020684
15	1168	15	2849	16	0,99	0,99	0,99	0,99	0,98	0,9931	0,162615	0,9923	0,015930
16	1182	24	2836	6	0,98	0,99	0,99	0,99	0,98	0,9940	0,175197	0,9926	0,016468
17	1155	17	2860	16	0,99	0,99	0,99	0,99	0,98	0,9931	0,158550	0,9918	0,015696
18	1140	14	2883	11	0,99	0,99	1,00	0,99	0,98	0,9933	0,168628	0,9938	0,017384
19	1126	12	2896	14	0,99	0,99	1,00	0,99	0,98	0,9922	0,163860	0,9936	0,019769
20	1102	18	2916	12	0,98	0,99	0,99	0,99	0,98	0,9932	0,154985	0,9926	0,015617
21	1191	11	2876	20	0,99	0,98	1,00	0,99	0,98	0,9932	0,169287	0,9923	0,017495
22	1155	21	2861	11	0,98	0,99	0,99	0,99	0,98	0,9931	0,187094	0,9921	0,018042
23	1170	11	2852	15	0,99	0,99	1,00	0,99	0,98	0,9931	0,154615	0,9936	0,015874
24	1194	16	2822	16	0,99	0,99	0,99	0,99	0,98	0,9940	0,150879	0,9921	0,020938
25	1151	12	2871	14	0,99	0,99	1,00	0,99	0,98	0,9931	0,184213	0,9936	0,018117
26	1123	15	2903	7	0,99	0,99	0,99	0,99	0,99	0,9927	0,182463	0,9946	0,018466
27	1156	12	2861	19	0,99	0,98	1,00	0,99	0,98	0,9936	0,150413	0,9923	0,015130
28	1163	9	2867	9	0,99	0,99	1,00	0,99	0,99	0,9922	0,162620	0,9956	0,018167
29	1141	18	2880	9	0,98	0,99	0,99	0,99	0,98	0,9940	0,175411	0,9933	0,017930
30	1187	9	2830	22	0,99	0,98	1,00	0,99	0,98	0,9915	0,157923	0,9923	0,015576
31	1162	12	2856	18	0,99	0,98	1,00	0,99	0,98	0,9933	0,165615	0,9926	0,016129
32	1165	17	2850	16	0,99	0,99	0,99	0,99	0,98	0,9932	0,168451	0,9918	0,019077
33	1123	18	2888	19	0,98	0,98	0,99	0,98	0,98	0,9946	0,158934	0,9909	0,017568
34	1148	17	2873	10	0,99	0,99	0,99	0,99	0,98	0,9936	0,174164	0,9933	0,018328
35	1136	10	2889	13	0,99	0,99	1,00	0,99	0,99	0,9914	0,148988	0,9943	0,015776
36	1129	10	2895	14	0,99	0,99	1,00	0,99	0,99	0,9915	0,190374	0,9941	0,020715
37	1161	15	2857	15	0,99	0,99	0,99	0,99	0,98	0,9924	0,163444	0,9926	0,016638
38	1182	15	2838	13	0,99	0,99	0,99	0,99	0,98	0,9927	0,156315	0,9931	0,015679
39	1149	18	2869	12	0,98	0,99	0,99	0,99	0,98	0,9943	0,177044	0,9926	0,018906
40	1161	13	2861	13	0,99	0,99	1,00	0,99	0,98	0,9922	0,159834	0,9936	0,017891
41	1173	19	2850	6	0,98	0,99	0,99	0,99	0,99	0,9926	0,175695	0,9938	0,017073
42	1151	11	2864	22	0,99	0,98	1,00	0,99	0,98	0,9928	0,145404	0,9918	0,016485
43	1187	11	2837	13	0,99	0,99	1,00	0,99	0,99	0,9935	0,173651	0,9941	0,015778
44	1159	16	2856	17	0,99	0,99	0,99	0,99	0,98	0,9927	0,187113	0,9918	0,018134
45	1153	19	2861	15	0,98	0,99	0,99	0,99	0,98	0,9943	0,190592	0,9916	0,021142
46	1163	11	2859	15	0,99	0,99	1,00	0,99	0,98	0,9933	0,165566	0,9936	0,017691
47	1150	20	2864	14	0,98	0,99	0,99	0,99	0,98	0,9942	0,174187	0,9916	0,016939
48	1170	16	2859	3	0,99	1,00	0,99	0,99	0,99	0,9930	0,157859	0,9953	0,018671
49	1148	13	2873	14	0,99	0,99	1,00	0,99	0,98	0,9930	0,167332	0,9933	0,017707

Tabela 11 – Execuções do modelo utilizando *Gaussian Naive Bayes*.

	Matriz de confusão				Métricas de desempenho					Desempenho do treinamento		Desempenho deo teste	
Execução	TP	FP	TN	FN	Precisão	Sensibilidade	Especificidade	<i>F-score</i>	MCC	Acurácia	Tempo de execução (s)	Acurácia	Tempo médio de execução (s)
0	1106	79	2862	1	0,93	1,00	0,97	0,97	0,95	0,9822	0,077000	0,9802	0,006792
1	1072	46	2921	9	0,96	0,99	0,98	0,97	0,97	0,9851	0,093669	0,9864	0,002635
2	1088	71	2887	2	0,94	1,00	0,98	0,97	0,96	0,9821	0,043719	0,9820	0,002696
3	1093	85	2869	1	0,93	1,00	0,97	0,96	0,95	0,9838	0,026656	0,9788	0,002567
4	1056	68	2923	1	0,94	1,00	0,98	0,97	0,96	0,9814	0,033986	0,9830	0,002599
5	1083	86	2876	3	0,93	1,00	0,97	0,96	0,95	0,9835	0,046587	0,9780	0,003698
6	1124	75	2849	0	0,94	1,00	0,97	0,97	0,96	0,9810	0,035755	0,9815	0,002564
7	1082	94	2872	0	0,92	1,00	0,97	0,96	0,94	0,9830	0,034943	0,9768	0,002922
8	1090	74	2883	1	0,94	1,00	0,97	0,97	0,95	0,9812	0,036407	0,9815	0,004009
9	1062	83	2903	0	0,93	1,00	0,97	0,96	0,95	0,9830	0,041714	0,9795	0,005205
10	1091	60	2897	0	0,95	1,00	0,98	0,97	0,96	0,9799	0,039020	0,9852	0,003780
11	1031	85	2932	0	0,92	1,00	0,97	0,96	0,95	0,9813	0,034055	0,9790	0,002610
12	1096	74	2877	1	0,94	1,00	0,97	0,97	0,96	0,9820	0,028132	0,9815	0,002665
13	1146	69	2831	2	0,94	1,00	0,98	0,97	0,96	0,9813	0,027071	0,9825	0,002394
14	1086	81	2881	0	0,93	1,00	0,97	0,96	0,95	0,9826	0,029413	0,9800	0,002608
15	1106	71	2871	0	0,94	1,00	0,98	0,97	0,96	0,9808	0,031685	0,9825	0,002716
16	1067	77	2903	1	0,93	1,00	0,97	0,96	0,95	0,9815	0,029694	0,9807	0,002796
17	1132	81	2834	1	0,93	1,00	0,97	0,96	0,95	0,9823	0,028559	0,9797	0,002503
18	1135	85	2826	2	0,93	1,00	0,97	0,96	0,95	0,9833	0,028994	0,9785	0,002551
19	1110	58	2879	1	0,95	1,00	0,98	0,97	0,96	0,9794	0,030023	0,9854	0,003064
20	1064	71	2910	3	0,94	1,00	0,98	0,97	0,95	0,9815	0,028981	0,9817	0,002856
21	1105	69	2874	0	0,94	1,00	0,98	0,97	0,96	0,9815	0,030988	0,9830	0,002602
22	1129	67	2852	0	0,94	1,00	0,98	0,97	0,96	0,9803	0,029323	0,9834	0,002761
23	1091	97	2860	0	0,92	1,00	0,97	0,96	0,94	0,9844	0,028436	0,9760	0,003242
24	1044	69	2934	1	0,94	1,00	0,98	0,97	0,96	0,9813	0,030282	0,9827	0,002702
25	1097	78	2872	1	0,93	1,00	0,97	0,97	0,95	0,9809	0,036925	0,9805	0,004331
26	1113	74	2860	1	0,94	1,00	0,97	0,97	0,96	0,9837	0,027667	0,9815	0,002494
27	1084	85	2879	0	0,93	1,00	0,97	0,96	0,95	0,9826	0,029989	0,9790	0,003028
28	1152	78	2817	1	0,94	1,00	0,97	0,97	0,95	0,9823	0,027079	0,9805	0,002547
29	1061	63	2922	2	0,94	1,00	0,98	0,97	0,96	0,9801	0,026594	0,9839	0,002746
30	1114	59	2873	2	0,95	1,00	0,98	0,97	0,96	0,9817	0,027811	0,9849	0,002781
31	1055	58	2935	0	0,95	1,00	0,98	0,97	0,96	0,9808	0,027729	0,9857	0,002710
32	1116	72	2858	2	0,94	1,00	0,98	0,97	0,96	0,9813	0,030776	0,9817	0,002666
33	1111	62	2874	1	0,95	1,00	0,98	0,97	0,96	0,9809	0,029132	0,9844	0,002867
34	1094	70	2883	1	0,94	1,00	0,98	0,97	0,96	0,9823	0,029015	0,9825	0,002798
35	1095	78	2875	0	0,93	1,00	0,97	0,97	0,95	0,9813	0,029633	0,9807	0,002899
36	1128	70	2848	2	0,94	1,00	0,98	0,97	0,96	0,9820	0,030797	0,9822	0,002591
37	1092	79	2877	0	0,93	1,00	0,97	0,97	0,95	0,9807	0,028817	0,9805	0,002991
38	1083	72	2893	0	0,94	1,00	0,98	0,97	0,96	0,9821	0,027316	0,9822	0,002470
39	1122	75	2851	0	0,94	1,00	0,97	0,97	0,96	0,9814	0,029401	0,9815	0,002894
40	1077	70	2899	2	0,94	1,00	0,98	0,97	0,96	0,9819	0,028398	0,9822	0,002636
41	1090	93	2864	1	0,92	1,00	0,97	0,96	0,94	0,9839	0,027729	0,9768	0,002736
42	1077	68	2901	2	0,94	1,00	0,98	0,97	0,96	0,9810	0,027360	0,9827	0,002724
43	1075	73	2898	2	0,94	1,00	0,98	0,97	0,95	0,9824	0,029273	0,9815	0,002859
44	1144	75	2827	2	0,94	1,00	0,97	0,97	0,95	0,9826	0,026835	0,9810	0,002483
45	1095	81	2871	1	0,93	1,00	0,97	0,96	0,95	0,9819	0,027019	0,9797	0,002583
46	1082	74	2891	1	0,94	1,00	0,97	0,97	0,95	0,9822	0,031139	0,9815	0,002514
47	1082	68	2895	3	0,94	1,00	0,98	0,97	0,96	0,9822	0,026770	0,9825	0,002549
48	1087	78	2883	0	0,93	1,00	0,97	0,97	0,95	0,9807	0,031826	0,9807	0,002613
49	1092	74	2880	2	0,94	1,00	0,97	0,97	0,95	0,9830	0,026394	0,9812	0,002478

Tabela 12 – Execuções do modelo utilizando Máquinas de Vetores Suporte.

	Matriz de confusão				Métricas de desempenho					Desempenho do treinamento		Desempenho do teste	
Execução	TP	FP	TN	FN	Precisão	Sensibilidade	Especificidade	<i>F-score</i>	MCC	Acurácia	Tempo de execução (s)	Acurácia	Tempo médio de execução (s)
0	1166	31	2846	5	0,97	1,00	0,99	0,98	0,98	0,9908	54,781028	0,9911	5,442610
1	1153	29	2865	1	0,98	1,00	0,99	0,99	0,98	0,9905	49,133175	0,9926	4,085891
2	1145	33	2869	1	0,97	1,00	0,99	0,99	0,98	0,9911	33,745709	0,9916	4,189034
3	1131	32	2881	4	0,97	1,00	0,99	0,98	0,98	0,9911	32,357095	0,9911	4,112504
4	1108	33	2903	4	0,97	1,00	0,99	0,98	0,98	0,9908	34,086623	0,9909	4,203380
5	1092	36	2920	0	0,97	1,00	0,99	0,98	0,98	0,9911	33,246840	0,9911	4,405161
6	1147	39	2856	6	0,97	0,99	0,99	0,98	0,97	0,9911	32,232699	0,9889	4,003678
7	1154	39	2852	3	0,97	1,00	0,99	0,98	0,98	0,9914	32,552933	0,9896	4,068027
8	1149	28	2868	3	0,98	1,00	0,99	0,99	0,98	0,9910	32,773149	0,9923	4,329439
9	1199	41	2799	9	0,97	0,99	0,99	0,98	0,97	0,9920	31,155111	0,9876	3,947725
10	1174	34	2839	1	0,97	1,00	0,99	0,99	0,98	0,9906	32,867285	0,9914	4,131800
11	1146	39	2860	3	0,97	1,00	0,99	0,98	0,97	0,9908	31,457840	0,9896	4,050183
12	1132	31	2882	3	0,97	1,00	0,99	0,99	0,98	0,9913	32,703665	0,9916	4,520870
13	1164	36	2846	2	0,97	1,00	0,99	0,98	0,98	0,9908	32,075855	0,9906	4,243501
14	1153	39	2855	1	0,97	1,00	0,99	0,98	0,98	0,9907	31,879488	0,9901	3,985429
15	1141	30	2873	4	0,97	1,00	0,99	0,99	0,98	0,9905	32,688911	0,9916	4,063804
16	1136	34	2875	3	0,97	1,00	0,99	0,98	0,98	0,9909	31,935144	0,9909	4,238563
17	1114	38	2886	10	0,97	0,99	0,99	0,98	0,97	0,9915	32,619445	0,9881	4,062493
18	1142	28	2876	2	0,98	1,00	0,99	0,99	0,98	0,9908	32,983480	0,9926	4,242272
19	1132	32	2881	3	0,97	1,00	0,99	0,98	0,98	0,9909	31,984025	0,9914	4,041250
20	1113	32	2895	8	0,97	0,99	0,99	0,98	0,98	0,9899	32,787182	0,9901	4,116491
21	1111	39	2898	0	0,97	1,00	0,99	0,98	0,98	0,9905	32,462263	0,9904	4,298876
22	1121	33	2889	5	0,97	1,00	0,99	0,98	0,98	0,9911	36,018703	0,9906	4,057200
23	1116	39	2891	2	0,97	1,00	0,99	0,98	0,98	0,9907	32,867412	0,9899	4,085330
24	1163	32	2853	0	0,97	1,00	0,99	0,99	0,98	0,9899	31,710740	0,9921	4,118508
25	1156	27	2862	3	0,98	1,00	0,99	0,99	0,98	0,9903	32,528358	0,9926	4,064336
26	1135	34	2875	4	0,97	1,00	0,99	0,98	0,98	0,9912	32,864842	0,9906	4,088093
27	1114	32	2900	2	0,97	1,00	0,99	0,98	0,98	0,9907	32,416743	0,9916	4,165167
28	1158	33	2856	1	0,97	1,00	0,99	0,99	0,98	0,9911	32,405593	0,9916	4,110906
29	1170	35	2839	4	0,97	1,00	0,99	0,98	0,98	0,9916	31,398000	0,9904	4,178521
30	1130	36	2880	2	0,97	1,00	0,99	0,98	0,98	0,9908	33,069159	0,9906	4,073953
31	1112	30	2904	2	0,97	1,00	0,99	0,99	0,98	0,9910	33,628299	0,9921	4,134573
32	1147	33	2864	4	0,97	1,00	0,99	0,98	0,98	0,9911	32,006229	0,9909	4,281817
33	1126	35	2884	3	0,97	1,00	0,99	0,98	0,98	0,9911	32,841846	0,9906	4,039294
34	1184	35	2825	4	0,97	1,00	0,99	0,98	0,98	0,9902	33,524435	0,9904	4,135319
35	1131	38	2876	3	0,97	1,00	0,99	0,98	0,98	0,9912	32,387064	0,9899	4,060697
36	1081	34	2930	3	0,97	1,00	0,99	0,98	0,98	0,9911	33,495234	0,9909	4,183260
37	1133	40	2873	2	0,97	1,00	0,99	0,98	0,97	0,9905	32,839800	0,9896	4,291011
38	1143	23	2879	3	0,98	1,00	0,99	0,99	0,98	0,9889	32,739242	0,9936	4,072031
39	1145	36	2865	2	0,97	1,00	0,99	0,98	0,98	0,9906	32,905180	0,9906	4,101672
40	1150	26	2866	6	0,98	0,99	0,99	0,99	0,98	0,9906	32,093870	0,9921	4,274784
41	1128	38	2881	1	0,97	1,00	0,99	0,98	0,98	0,9917	33,009707	0,9904	4,112308
42	1162	36	2843	7	0,97	0,99	0,99	0,98	0,97	0,9915	33,880344	0,9894	4,087525
43	1129	28	2889	2	0,98	1,00	0,99	0,99	0,98	0,9896	32,377294	0,9926	4,112854
44	1118	37	2890	3	0,97	1,00	0,99	0,98	0,98	0,9911	32,515887	0,9901	4,058528
45	1152	38	2857	1	0,97	1,00	0,99	0,98	0,98	0,9910	32,147313	0,9904	4,218342
46	1075	39	2931	3	0,96	1,00	0,99	0,98	0,97	0,9917	32,531404	0,9896	4,136688
47	1112	33	2900	3	0,97	1,00	0,99	0,98	0,98	0,9905	33,400336	0,9911	4,149040
48	1136	39	2868	5	0,97	1,00	0,99	0,98	0,97	0,9910	32,428194	0,9891	4,146903
49	1153	38	2852	5	0,97	1,00	0,99	0,98	0,97	0,9916	31,986236	0,9894	4,012471