UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

A utilização de algoritmos de aprendizado de máquina em problemas de classificação

Maria Rita Sifuentes Batista

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Maria Rita Sifuentes Batista

A utilização de algoritmos de aprendizado de máquina em problemas de classificação

Dissertação apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP,
como parte dos requisitos para obtenção do título
de Mestra – Mestrado Profissional em Matemática,
Estatística e Computação Aplicadas à Indústria.
VERSÃO REVISADA

Área de Concentração: Matemática, Estatística e
Computação

Orientador: Prof. Dr. Fabrício Simeoni de Sousa

USP – São Carlos
Janeiro de 2019

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

B333u Batista, Maria Rita Sifuentes
 A utilização de algoritmos de aprendizado de
 máquina em problemas de classificação / Maria Rita
 Sifuentes Batista; orientador Fabrício Simeoni de
 Sousa. -- São Carlos, 2019.
 109 p.

 Dissertação (Mestrado - Programa de Pós-Graduação
 em Mestrado Profissional em Matemática, Estatística
 e Computação Aplicadas à Indústria) -- Instituto de
 Ciências Matemáticas e de Computação, Universidade
 de São Paulo, 2019.

 1. Previsão. 2. Aprendizado de Máquina. 3.
 Classificação. 4. Variáveis. 5. Algoritmo. I. Sousa,
 Fabrício Simeoni de, orient. II. Título.

Maria Rita Sifuentes Batista

The use of machine learning algorithms in classification problems

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master – Professional Masters in Mathematics, Statistics and Computing Applied to Industry. *FINAL VERSION*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. Fabrício Simeoni de Sousa

USP – São Carlos
January 2019

“Não desistir é indispensável.”
(Stephen Hawking)

RESUMO

BATISTA, M. R. S. **A utilização de algoritmos de aprendizado de máquina em problemas de classificação**. 2019. 109 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Os últimos anos foram marcados por um avanço expressivo da tecnologia, principalmente na área de computação. Estes avanços, quando somados à diversidade de produtos oferecidos por empresas de diferentes segmentos, e aos esforços destas em capturar e armazenar dados de seus clientes e de suas operações, ajudam a explicar a quantidade de informações que atualmente é produzida. As organizações, em geral, têm se mostrado eficientes em capturar, organizar e armazenar grandes quantidades de dados, mas nem todas os utilizam adequadamente, no sentido de transformá-los em conhecimentos úteis para suas atividades. Algoritmos de aprendizado de máquina são uma ferramenta computacional poderosa para aquisição de conhecimento a partir da experiência. A utilização desses algoritmos permite avanços e descobertas que conferem vantagem competitiva às empresas. A tarefa de aprendizado de máquina mais comum é o aprendizado supervisionado, cujo objetivo é aprender um modelo preditivo a partir de um conjunto de dados. Esse modelo deve ser capaz de generalizar o conhecimento adquirido para dados desconhecidos. Isso permite que o modelo tenha uma boa capacidade preditiva. Uma aplicação importante e bastante utilizada do aprendizado supervisionado são os problemas de classificação, comumente encontrados na indústria financeira. Um dos desafios dessa indústria é prever a capacidade de pagamento de seus clientes, classificando-os como bons ou maus pagadores. Neste trabalho, cinco algoritmos de aprendizado de máquina supervisionado foram investigados e aplicados à um problema real de classificação: regressão logística, classificadores bayesianos, k-vizinhos mais próximos, *random forests* e redes neurais. Como o desempenho desses algoritmos é afetado pelas variáveis utilizadas, técnicas de seleção de variáveis foram aplicadas ao conjunto de dados original. O uso dessas técnicas permite reduzir o tempo computacional, removendo informações redundantes e irrelevantes. Medidas de desempenho para classificação binária foram utilizadas para avaliar o desempenho preditivo dos modelos gerados pelos cinco algoritmos e compará-los. Como é cada vez mais importante ter modelos facilmente interpretáveis, foram também avaliadas a interpretabilidade e a complexidade dos modelos gerados.

Palavras-chave: Previsão, Aprendizado de Máquina, Classificação, Variáveis, Algoritmo.

ABSTRACT

BATISTA, M. R. S. **The use of machine learning algorithms in classification problems**. 2019. 109 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

The last few years were remarkable by relevant advances in technology, mainly related to computers. These advances, when added to the diversity of products offered by companies from different segments and their efforts in capturing and storing data from their customers and operations, helps to explain the amount of information that is currently being produced. Overall, the organizations have been efficient in capturing, organizing, and storing large amounts of data, but not all of them uses it adequately to make them useful knowledge for their activities. Learning algorithms are a powerful machine toll to acquire knowledge based on experience. The use of these algorithms allows advances and discoveries that brings a competitive advantage to the companies. The most common machine learning task is supervised learning, whose objective is to learn a predictive model from a set of data. This model should be able to generalize the acquired knowledge to a set of unknown data. This allows the model to have a good predictive capability. An important and widely used application of supervised learning are the classification problems, commonly seen in the financial industry. One of the challenges of this industry is to predict the payment capacity of its customers, rating them as good or bad payers. In this study, five supervised machine learning algorithms, logistic regression, Bayesian classifiers, k-neighbors, random forests and neural networks were investigated and applied to a real classification problem. Since the performance of these algorithms are affected by the variables used, variable selection techniques were applied to the original data set. The use of these techniques allows a computational reduction time by removing redundant and irrelevant information. Performance measures for binary classification were used to evaluate the predictive performance of the models generated by the five algorithms and to compare them. Since it is increasing the importance to have easily interpretable models, the interpretability and complexity of the models generated were also evaluated.

Keywords: Prediction, Machine Learning, Classification, Variables, Algorithm.

LISTA DE ILUSTRAÇÕES

Figura 1 – Distribuições da variável resposta e das variáveis categóricas sexo, escolaridade e estado civil dentre as classes da variável resposta.	32
Figura 2 – Distribuição da variável atraso máximo.	33
Figura 3 – Distribuição da variável atraso máximo condicionada à variável resposta. . .	34
Figura 4 – Histograma da variável Limite em escala linear (ao lado esquerdo) e em escala logarítmica (ao lado direito).	35
Figura 5 – Histograma da variável Idade em escala linear (ao lado esquerdo) e em escala logarítmica (ao lado direito).	35
Figura 6 – Histograma da variável Saldo Acumulado Médio em escala linear (ao lado esquerdo) e em escala logarítmica (ao lado direito).	36
Figura 7 – <i>Boxplot</i> das variáveis Soma de Saldo Devedor e Saldo Acumulado Médio, por classe da variável resposta.	36
Figura 8 – <i>Boxplot</i> das variáveis Limite e <i>Limit Ratio</i> , por classe da variável resposta. .	37
Figura 9 – <i>Boxplot</i> das variáveis Limite e Saldo Acumulado Médio por categorias das variáveis Sexo e Idade.	38
Figura 10 – Possíveis associações lineares entre duas variáveis quantitativas e os respectivos valores do coeficiente de Pearson.	40
Figura 11 – Exemplos de relacionamentos monotônicos e não-monotônicos entre duas variáveis.	41
Figura 12 – Matriz das correlações de Pearson das variáveis quantitativas contínuas. . .	43
Figura 13 – Matriz das correlações de Spearman das variáveis qualitativas ordinais. . .	44
Figura 14 – Gráfico de dispersão das variáveis quantitativas contínuas.	45
Figura 15 – Exemplo de classificação do método KNN com dois rótulos de classes e $k=7$. .	61
Figura 16 – Árvore de decisão (à esquerda) e as regiões de decisão no espaço de atributos x^1 e x^2	67
Figura 17 – Comparação das medidas de impureza para problemas de classificação binária. .	68
Figura 18 – Modelo matemático de um neurônio proposto por McCulloch e Pitts (1943). .	72
Figura 19 – Rede perceptron.	74
Figura 20 – Conjunto de exemplos de treinamento separáveis linearmente (à esquerda) e não separáveis linearmente (à direita).	75
Figura 21 – Rede neural consistindo de múltiplas camadas de neurônios conectados (MLP). .	76
Figura 22 – Matriz de confusão para um problema de classificação de duas classes. . . .	85
Figura 23 – Gráfico ROC e área sob a curva ROC para dois classificadores distintos. . .	87

Figura 24 – Distribuição da variável categórica status incondicionada à variável resposta (à esquerda) e condicionada à variável resposta (à direita) após a categorização.	90
Figura 25 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para determinação do ponto de corte para o método Naive Bayes.	92
Figura 26 – Taxa de erro média obtida através de validação cruzada com $k = 10$ para determinação do valor do hiperparâmetro número de vizinhos.	94
Figura 27 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para determinação do ponto de corte para o método KNN.	95
Figura 28 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para determinação do ponto de corte para o método Regressão Logística.	96
Figura 29 – Taxa de erro OOB de acordo com o número de árvores na floresta aleatória.	97
Figura 30 – Taxa de erro OOB média obtida através de validação cruzada com $k = 10$ para determinação do valor do hiperparâmetro <i>mtry</i>	98
Figura 31 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para determinação do ponto de corte para o método <i>Random Forests</i>	99
Figura 32 – Taxa de erro média obtida através de validação cruzada com $k = 10$ para determinação do número de camadas ocultas e do número de neurônios em cada camada.	100
Figura 33 – Taxa de erro média obtida através de validação cruzada com $k = 10$ para determinação do número de camadas ocultas e do número de neurônios em cada camada.	101
Figura 34 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para determinação do ponto de corte para o método Redes Neurais.	102
Figura 35 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para comparação dos modelos de classificação.	104

LISTA DE TABELAS

Tabela 1 – Especificação das variáveis.	24
Tabela 2 – Categorização da variável Idade.	25
Tabela 3 – Interpretação do relacionamento linear de Pearson.	40
Tabela 4 – Distribuição conjunta dos valores absolutos e relativos dos clientes de acordo com o sexo e as categorias da variável resposta.	49
Tabela 5 – Valores esperados na tabela Tabela 4 supondo independência entre as variáveis.	50
Tabela 6 – Estatística qui-quadrado de Pearson e coeficiente de correlação entre as variáveis qualitativas e a variável resposta.	51
Tabela 7 – Significância do teste qui-quadrado para linearidade aplicado às variáveis qualitativas ordinais e a variável resposta.	53
Tabela 8 – Fisher <i>Ratio</i> das variáveis quantitativas contínuas.	54
Tabela 9 – Correlações <i>Point Biserial</i> entre a variável resposta e as variáveis quantitativas e significância do teste.	56
Tabela 10 – Medidas de Associação recomendadas de acordo com os tipo de variáveis envolvidas na análise.	56
Tabela 11 – Relação entre probabilidade e <i>odds</i>	63
Tabela 12 – Ponto de corte baseado na especificidade e sensibilidade para cada amostra k da validação cruzada.	92
Tabela 13 – Medidas de desempenho do método Naive Bayes obtidas a partir do conjunto de dados de teste.	93
Tabela 14 – Ponto de corte baseado na especificidade e sensibilidade para cada amostra k da validação cruzada.	94
Tabela 15 – Medidas de desempenho do método KNN obtidas a partir do conjunto de dados de teste.	95
Tabela 16 – Ponto de corte baseado na especificidade e sensibilidade para cada amostra k da validação cruzada.	96
Tabela 17 – Medidas de desempenho do método regressão logística obtidas a partir do conjunto de dados de teste.	97
Tabela 18 – Ponto de corte baseado na especificidade e sensibilidade para cada amostra k da validação cruzada.	98
Tabela 19 – Medidas de desempenho do método <i>random forests</i> obtidas a partir do conjunto de dados de teste.	99

Tabela 20 – Ponto de corte baseado na especificidade e sensibilidade para cada amostra k da validação cruzada.	101
Tabela 21 – Medidas de desempenho do método redes neurais obtidas a partir do conjunto de dados de teste.	102
Tabela 22 – Medidas de desempenho dos métodos de classificação.	103

SUMÁRIO

1	INTRODUÇÃO	17
1.1	Conceitos Iniciais	17
1.1.1	<i>Aprendizado de Máquina e o Problema de Classificação</i>	17
1.1.2	<i>Variáveis e Conjuntos de Dados</i>	21
1.2	Definição do Problema e Descrição dos Dados Utilizados	22
2	ANÁLISE EXPLORATÓRIA	27
2.1	Análise Univariada	27
2.2	Análise Multivariada	38
3	REDUÇÃO DE DIMENSIONALIDADE	47
3.1	Técnicas de Redução de Dimensionalidade	47
3.1.1	<i>Seleção de Variáveis</i>	48
3.1.1.1	<i>Métodos de Seleção de Variáveis Baseados em Filtros</i>	49
4	MÉTODOS DE CLASSIFICAÇÃO	57
4.1	Classificador Naive Bayes	57
4.2	K-Nearest Neighbor	59
4.3	Regressão Logística	61
4.4	Random Forests	66
4.5	Redes Neurais	71
5	AVALIAÇÃO DOS MODELOS PREDITIVOS	83
5.1	Amostragem	83
5.2	Métricas de Desempenho para Problemas de Duas Classes	84
6	RESULTADOS	89
6.1	Tratamentos Gerais dos Dados	89
6.2	Desempenho dos Classificadores	91
6.3	Comparativo de Desempenho	102
7	CONCLUSÃO	105
	REFERÊNCIAS	107

INTRODUÇÃO

Este capítulo apresenta a revisão bibliográfica sobre o tema aprendizado de máquina e define, com o devido embasamento teórico, o problema de classificação. Também apresenta a caracterização de um conjunto de dados, passando pelo conceito de variável e fazendo a adequada distinção entre variáveis explicativas e variável resposta, além de evidenciar outras definições que serão utilizadas ao longo do texto. A base de dados real utilizada no estudo também é apresentada, o que inclui especificações das variáveis originais e das que foram criadas com o objetivo de ampliar o conjunto de variáveis preditoras.

1.1 Conceitos Iniciais

1.1.1 *Aprendizado de Máquina e o Problema de Classificação*

A teoria de aprendizado de máquina está pautada no conceito de aprendizagem e este aspecto é, além de notório, preliminarmente abordado na literatura desta disciplina científica. Diversos autores definem aprendizagem como a capacidade de um programa computacional melhorar o seu desempenho através da experiência. Diz-se que um programa de computador aprende a partir de uma experiência E em relação a uma certa classe T de tarefas e medida de desempenho P , se o seu desempenho nas tarefas T , tal como medido por P , melhora com a experiência E (MITCHELL, 1997). O aprendizado de máquina, para Marsland (2014), consiste em fazer com que os computadores modifiquem ou adaptem suas ações para que estas se tornem mais precisas. Luxburg e Schölkopf (2008), por sua vez, afirmam que aprendizagem refere-se ao processo de inferir regras gerais através da observação de exemplos, o que quer dizer que para que uma máquina ou computador aprenda tarefas específicas seguindo algoritmos de aprendizagem, à ela são revelados exemplos particulares desta tarefa e o objetivo é então inferir uma regra geral que pode tanto explicar os exemplos já vistos quanto generalizar para novos exemplos nunca antes vistos. De maneira similar, Marsland (2014) destaca que o conceito chave para as

máquinas é aprendizagem a partir dos dados e que a generalização, definida como a identificação de similaridade entre diferentes situações, é o que torna a aprendizagem útil, pois o conhecimento adquirido pode ser aplicado em diversos cenários. Para [Faceli et al. \(2011\)](#), a capacidade de aprendizado é considerada essencial para um comportamento inteligente, e atividades como memorizar, observar e explorar situações para aprender fatos estão relacionadas a aprendizagem.

O aprendizado de máquina tem raízes na inteligência artificial, estatística e ciência da computação, mas agora estabeleceu-se como uma disciplina por direito próprio ([LUXBURG; SCHÖLKOPF, 2008](#)). Neste sentido, [Marsland \(2014\)](#) também declara que o aprendizado de máquina mescla ideias da neurociência e da biologia, estatística, matemática e física, e só na última década a sua inerente multidisciplinaridade foi reconhecida. Ainda segundo estes autores, o aprendizado de máquina, diferente da inteligência artificial, não procura gerar um comportamento inteligente, mas descobrir mecanismos através dos quais tarefas específicas podem ser aprendidas por computadores - o foco é a inferência indutiva e a capacidade de generalização.

Em geral, os problemas de aprendizado podem ser classificados em:

- Supervisionados: um conjunto de exemplos com as respostas corretas é fornecido e, baseado neste conjunto, o algoritmo escolhido generaliza para responder corretamente à todas as possíveis futuras entradas. À este conjunto dá-se o nome de conjunto de treinamento, pois é a partir destes exemplos que o algoritmo será treinado para fazer generalizações. Algoritmos clássicos do aprendizado supervisionado incluem, por exemplo, análise de regressão linear, análise de regressão logística e abordagens mais modernas como *support vector machine*, *boosting*, dentre outras.
- Não supervisionados: respostas corretas para cada um dos objetos de entrada não são fornecidas e, portanto, os algoritmos procuram identificar similaridades entre os objetos de entrada de maneira que aqueles que apresentam algo em comum são categorizados conjuntamente. Pode-se buscar entender as relações entre as variáveis ou entre os objetos de entrada. Este cenário é dito não supervisionado dada a ausência de um vetor de respostas para supervisionar a análise. Uma ferramenta muito utilizada no aprendizado não supervisionado é a análise de *cluster*.

O problema de classificação é o mais estudado em aprendizado supervisionado e consiste em tomar vetores de entrada e decidir à qual das c classes eles pertencem com base nos exemplos de treinamento. Mais formalmente, seguindo a notação de [Luxburg e Schölkopf \(2008\)](#), em problemas de classificação dois tipos de espaço devem ser considerados: o espaço X das instâncias ou objetos e o espaço Y das saídas ou categorias. No caso de classificação binária, o espaço das saídas é identificado pelo conjunto $\{-1, +1\}$ ou $\{0, +1\}$. Segundo [Luxburg e Schölkopf \(2008\)](#), a fim de aprender, ao algoritmo são apresentados exemplos de treinamento com o rótulo

da categoria correspondente. O objetivo é encontrar um mapeamento $f : X \rightarrow Y$, chamado de classificador, que faz com que o número de objetos atribuídos equivocadamente à uma das categorias seja o menor possível. Um algoritmo de classificação, portanto, é um procedimento que toma como entrada os exemplos de treinamento e produz como saída um classificador f . [Faceli et al. \(2011\)](#) destacam que um algoritmo de aprendizado de máquina aprende uma aproximação $\hat{f}$ que permite estimar o valor produzido pelo classificador f para novas entradas, ou seja, $y = \hat{f}(x)$ para qualquer par (x, y) , com $x \in X$, $y \in Y$. Um classificador também pode ser designado por função ou modelo.

Os algoritmos de classificação, segundo [James et al. \(2013\)](#), podem ser divididos em paramétricos e não-paramétricos. Métodos paramétricos fazem uma suposição sobre a forma funcional de f , o que simplifica o problema, pois normalmente é mais fácil estimar um conjunto de parâmetros do que ajustar uma função inteiramente arbitrária. A desvantagem de uma abordagem como esta, no entanto, é que o método escolhido geralmente não corresponde à forma verdadeira desconhecida de f . Métodos não paramétricos, pelo contrário, não assumem explicitamente uma forma específica para f . Ao invés disso, segundo os autores, buscam uma estimativa que seja o mais próxima possível dos pontos de dados, e que confere o potencial de ajustar uma gama de formas possíveis para f . No entanto, uma vez que o problema não é reduzido a estimar um conjunto definido de parâmetros, um número grande de observações é necessário para obter uma estimativa precisa.

Segundo [James et al. \(2013\)](#), o objetivo de determinar uma função que relacione os objetos de entrada às suas saídas pode estar associado a dois motivos principais: (1) prever as respostas para futuros objetos, tarefa conhecida como predição ou (2) melhor compreender o relacionamento entre os objetos de entrada e saída, tarefa conhecida como inferência. Dependendo se o objetivo final é previsão, inferência ou uma combinação dos dois, diferentes métodos para estimar f podem ser apropriados. Em aprendizado de máquina supervisionado o objetivo é fazer previsão. Neste cenário, f é frequentemente tratada como uma caixa-preta, no sentido de que não se trata de conhecer a forma exata da função desde que ela produza boas previsões. Supondo que o conjunto de pares $\{(x_1, y_1), \dots, (x_n, y_n)\}$ tenha sido utilizado para obter a estimativa $\hat{f}$ do mapeamento f , em problemas em que se deseja fazer previsão o maior interesse não é saber se $\hat{f}(x_i) \approx y_i$, mas se $\hat{f}(x_0) \approx y_0$, em que (x_0, y_0) é um objeto de teste inédito, não utilizado para treinar o algoritmo de aprendizado. Existe, por isso, um outro conjunto de exemplos, chamado conjunto de teste e composto pelos objetos nunca vistos, que é utilizado para avaliar o desempenho de modelos preditivos. O único problema com isto é que se reduz a quantidade de dados disponível para treinamento.

Além da avaliação do desempenho do classificador, [Marsland \(2014\)](#) também coloca a necessidade de avaliar a suficiência do treinamento realizado. Se o treinamento ocorrer por muito tempo, então o algoritmo se adapta aos dados, o que significa que os ruídos e imprecisões também serão aprendidos e o modelo não será capaz de generalizar. À medida que o algoritmo

continua a aprender, acaba produzindo um modelo muito mais complexo. Este fenômeno é conhecido como *overfitting* e o processo de aprendizado precisa ser interrompido antes que ele ocorra. Um terceiro conjunto de dados, portanto, é necessário para validar o aprendizado, ou seja, para saber se o algoritmo está generalizando a cada etapa. Três conjuntos, ao final, são necessários: treinamento, validação e teste. Existem diversas maneiras de particionar o conjunto de exemplos original, e as principais estão descritas na [Seção 5.1](#).

Os tópicos abaixo, direcionados por [Marshall \(2014\)](#), descrevem o processo pelo qual os algoritmos de aprendizado de máquina são selecionados, aplicados e avaliados para o problema:

1. **Coleta e preparação dos dados:** os dados nem sempre estão imediatamente disponíveis para uso. Quando o problema é novo, por exemplo, os dados precisam ser coletados do início e preparados. A quantidade de dados, neste caso, precisa ser considerada, pois algoritmos de aprendizado de máquina requerem quantidades significativas de dados;
2. **Seleção de variáveis:** esta etapa consiste na identificação de variáveis que são de fato úteis para o problema em análise, o que requer conhecimento prévio do problema e dos dados e/ou emprego de técnicas de redução de dimensionalidade;
3. **Escolha do algoritmo:** definidos o problema e o conjunto de dados, um ou mais algoritmos de aprendizado de máquina são selecionados;
4. **Determinação dos parâmetros:** muitos algoritmos exigem que alguns parâmetros sejam inputados manualmente (por exemplo: número de iterações, tolerância ao erro, etc.), ou requerem experimentação prévia para identificação dos valores apropriados;
5. **Treinamento:** com os dados e os algoritmos disponíveis, a etapa do treinamento consiste no uso de recursos computacionais a fim de construir uma função f , ou um classificador, para prever as saídas em novos dados;
6. **Avaliação:** esta etapa inclui a seleção de métricas apropriadas para avaliação do desempenho do classificador construído. Para tanto, dados não utilizados no treinamento são necessários.

Na literatura é possível encontrar uma distinção entre parâmetros e hiperparâmetros de um modelo. Hiperparâmetros são parâmetros que definem a arquitetura de um método de aprendizado, sendo o processo de busca pela arquitetura ideal chamado de ajuste de hiperparâmetros - *hyperparameter tuning*. Enquanto os parâmetros do modelo especificam como transformar os objetos de entrada na saída desejada, os hiperparâmetros definem como o modelo é realmente estruturado. Os parâmetros são aprendidos durante o treinamento e hiperparâmetros, pelo contrário, não podem ser estimados diretamente do conjunto de treinamento. Como exemplificam [Wang et al. \(2018\)](#), em redes neurais os pesos são os parâmetros ajustados em tempo de treinamento e os

hiperparâmetros são, dentre outros, o número de neurônios nas camadas ocultas. A definição dos valores dos hiperparâmetros normalmente é feita por experimentação.

1.1.2 Variáveis e Conjuntos de Dados

O material básico com que o pesquisador trabalha são as observações. Elas são representadas por um vetor de características de interesse que as descreve, as quais são também denominadas variáveis ou atributos. As variáveis são mensuradas em cada elemento da amostra ou população e são assim chamadas por representarem características que assumem valores diferentes de indivíduo para indivíduo. Quando a medição de uma determinada característica apresenta-se ausente para um ou mais indivíduos, seja por falha no processo de coleta de informações ou mesmo por impossibilidade de completá-la, dizemos que aquele valor é *missing*. Cada observação, também denominada objeto ou registro, corresponde a uma ocorrência dos dados e cada variável está associada a uma propriedade específica dos objetos, podendo esta propriedade retratar noções físicas ou abstratas. Uma coleção de observações assim constituídas compõe um conjunto de dados. As variáveis podem assumir valores numéricos e não numéricos, sendo classificadas em:

- **Qualitativas ou categóricas:** são aquelas para as quais uma medição numérica não é possível, pois representam qualidades. Mesmo apesar de uma ordenação dos valores fazer sentido, nenhuma operação aritmética pode ser aplicada. Os tipos podem ser:
 1. Nominal: os valores não apresentam ordem definida, são apenas nomes diferentes. Exemplo: sexo, estado civil, cor dos olhos, etc.
 2. Ordinal: os valores apresentam uma ordem definida e é possível determinar se um valor é maior, menor ou igual ao outro. Exemplo: nível de escolaridade, estágio de alguma doença, meses do ano, etc.
- **Quantitativas ou numéricas:** são aquelas para as quais uma medição numérica é possível. Os tipos podem ser:
 1. Discretas: são aquelas originadas de processos de contagem e que, portanto, só podem assumir valores inteiros. Exemplos: número de filhos, etc.
 2. Contínuas: são características mensuráveis, originadas de processos de medições e que, portanto, podem assumir qualquer valor em uma escala contínua (reta real). Exemplo: altura, peso, idade, pressão arterial, etc.

Conjuntos de dados apresentam atributos de entrada, também denominados variáveis explicativas, preditoras ou independentes, e podem apresentar um atributo alvo, também chamado de *target*, variável resposta ou variável dependente, que representa o fenômeno de interesse sobre

o qual se deseja fazer previsões. Segundo [Mitchell \(1997\)](#), tanto em problemas de regressão, em que o atributo alvo contém valores numéricos contínuos, quanto de classificação, em que o atributo alvo contém rótulos que identificam categorias ou classes às quais os objetos pertencem, os demais atributos são denominados atributos preditivos por poderem ser utilizados para prever o valor do atributo alvo.

Equivalente à notação adotada por [Faceli et al. \(2011\)](#), conjuntos de dados podem ser representados por uma matriz de objetos $X_{n \times d}$, em que n é o número de objetos e d é o número de atributos de entrada. O valor d define a dimensão da base de dados e cada elemento x_i^j da matriz contém o valor da j -ésima característica para o i -ésimo objeto, com $j = 1, \dots, d$ e $i = 1, \dots, n$. Os elementos x_i^j também poderão ser designados por x_{ij} . A j -ésima variável preditora pode ser representada simplesmente por x^j e a i -ésima observação por x_i . De maneira correspondente, y_i é a variável resposta da i -ésima observação e y_c faz referência à uma das classes da variável resposta, $c = 1, \dots, l$. Um conjunto de treinamento, portanto, consiste de $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, em que $x_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$. Para fazer referência à um objeto qualquer da base de dados, x_i poderá ser designado apenas por x . Em alguns trechos do texto, o índice i será utilizado para fazer referência à uma variável e o índice k para fazer referência à uma observação do conjunto de dados.

1.2 Definição do Problema e Descrição dos Dados Utilizados

A base de dados utilizada neste trabalho contempla informações relacionadas ao comportamento de pagamento de clientes de um banco em Taiwan e pode ser obtida no endereço eletrônico do UCI Machine Learning Repository. No total, são 23 variáveis explicativas e 1 variável resposta binária, intitulada "Default Payment Next Month", que indica se foi devidamente realizado o pagamento da fatura do mês imediatamente seguinte ao último mês do histórico de dados coletados. Os dados disponibilizados capturam um período no qual os emissores de cartão de crédito em Taiwan enfrentaram a crise da dívida em dinheiro e limite de cartão de crédito, o que fez com que a inadimplência aumentasse. Segundo [Chou \(2006 apud YEH; LIEN, 2009\)](#), com o intuito de aumentar a participação de mercado, os bancos emissores taiwaneses superestimaram a concessão de dinheiro em espécie e cartões de crédito para candidatos não qualificados e, ao mesmo tempo, a maioria dos portadores de cartões, independentemente de sua capacidade de pagamento, utilizou excessivamente o cartão de crédito para consumo e acumulou dívidas pesadas, fatores que juntos motivaram a crise. A descrição detalhada das variáveis do conjunto de dados utilizado é apresentada abaixo:

- x^1 (quantitativa contínua) – Limite do cartão de crédito, em moeda local (NT dollar). Inclui o limite individual do cliente e limite adicional para dependentes;

- x^2 (qualitativa nominal) – Indica o gênero do cliente (Masculino=1, Feminino=2);
- x^3 (qualitativa ordinal) – Grau de escolaridade (Pós-Graduação=1, Graduação=2, Ensino Médio=3, Outros=4);
- x^4 (qualitativa ordinal) – Estado Civil (Casado(a)=1, Solteiro(a)=2, Outros=3);
- x^5 (quantitativa contínua) – Idade;
- $x^6 - x^{11}$ (qualitativa nominal) – Histórico de 6 meses, de Abril a Setembro de 2005, do *status* de pagamento mensal, sendo o índice de maior número referente ao mês mais remoto. Isto é, x^6 representa o *status* do pagamento em Setembro de 2005, x^7 o *status* do pagamento em Agosto de 2005 e assim por diante, até o *status* do pagamento em Abril de 2005, representado por x^{11} . A escala de medição para o *status* de pagamento é a seguinte: -2 indica que não houve consumo, ou seja, o limite do cartão não foi utilizado; -1 indica que a fatura foi paga integralmente; 0 indica que foi feito um pagamento entre o valor mínimo e o total da fatura, o que significa que o crédito rotativo foi acionado; 1 indica que o pagamento foi acusado com um mês de atraso; 2 indica que o pagamento da fatura foi acusado com dois meses de atraso; e assim sucessivamente, até o nono mês;
- $x^{12} - x^{17}$ (quantitativa contínua) – Histórico de 6 meses, de Abril a Setembro de 2005, do saldo mensal da fatura, sendo o índice de maior número referente ao mês mais remoto. Isto é, x^{12} representa o saldo da fatura em Setembro de 2005, x^{13} o saldo da fatura em Agosto de 2005 e assim por diante, até x^{17} , saldo da fatura em Abril de 2005;
- $x^{18} - x^{23}$ (quantitativa contínua) – Histórico de 6 meses, de Abril a Setembro de 2005, dos valores mensais efetivamente pagos. Da mesma maneira, x^{18} representa o valor pago em Setembro de 2005 e, por último, x^{23} representando o valor pago em Abril de 2005;
- x^{24} (qualitativa nominal) – Variável resposta binária, em que 0 sinaliza que o pagamento foi realizado e 1 sinaliza que o pagamento não foi realizado no mês avaliado, ou seja, o cliente é marcado como inadimplente.

As variáveis x^2 , x^3 e x^4 , originalmente quantitativas, foram transformadas em qualitativas atribuindo-se a cada valor numérico possível as categorias correspondentes. Os valores não documentados destas variáveis, ou seja, não explicitados na descrição acima, foram excluídos da análise.

Para simplificar a notação, foi feita uma associação entre os meses de referência das variáveis explicativas x^6 - x^{23} e números de 1 a 6. Assim, o número 1 faz referência a Setembro de 2005, o número 2 a Agosto de 2005 e, por último, o número 6 a Abril de 2005. Estas variáveis, portanto, podem ser representadas como $(\text{variável}_k)_{k=1}^6$.

A variável x^{24} , originalmente intitulada de *Default Payment Next Month*, poderá ser designada simplesmente por *Default*. É importante frisar que o conceito de *default* no âmbito do

gerenciamento do Risco de Crédito está atrelado à quantidade de dias de atraso registrada para o pagamento de uma obrigação de crédito¹. Isto significa que a operação não será marcada em descumprimento desde que esteja dentro de um intervalo pré-determinado de dias de atraso. No contexto dos dados utilizados neste trabalho, no entanto, o termo *default* é empregado no sentido literal, de negligência ou omissão, e será usado para fazer referência aos pagamentos de fatura não realizados no mês avaliado, independentemente da situação da operação nos meses anteriores. Isto significa que é possível encontrar clientes cuja variável $(STATUS_k)_{k=1}^6$ assume valores iguais a -2 ou -1 em seus seis pontos de aferição e o atributo alvo assume valor igual a 1.

As demais variáveis originais também foram renomeadas, como mostra a [Tabela 1](#):

Tabela 1 – Especificação das variáveis.

Variável Original	Variável Renomeada
x^1	Limite
x^2	Sexo.str
x^3	Escolaridade.str
x^4	Estado_Civil.str
x^5	Idade
$x^6 - x^{11}$	STATUS_1 - STATUS_6
$x^{12} - x^{17}$	FATURA_1 - FATURA_6
$x^{18} - x^{23}$	PGTO_1 - PGTO_6
x^{24}	Default

A fim de ampliar o conjunto de variáveis preditoras e, consequentemente, melhorar o desempenho dos métodos de classificação estudados, variáveis adicionais foram construídas a partir das variáveis originais. As descrições e formulações destas variáveis são apresentadas abaixo:

- Idade Categorizada (qualitativa ordinal): a variável idade foi categorizada e, para tanto, utilizou-se o método árvore de decisão ([Seção 4.4](#)) tomando como *target* a variável *Default* e como variável explicativa o próprio atributo idade. Nesta etapa o objetivo é simplesmente avaliar o comportamento da taxa de descumprimento em cada categoria delineada, ou seja, em cada faixa de idade. Sendo assim, a variável resposta é interpretada como sendo do tipo numérica discreta e a taxa de descumprimento média de cada categoria é dada pela média aritmética da variável *Default*, conforme mostra a [Tabela 2](#):
- Soma de Saldo Devedor (quantitativa contínua): soma das diferenças mensais das variáveis que representam o saldo da fatura e o valor efetivamente pago. Genericamente:

$$SOMA_SALDO_DEVEDOR = \sum_{k=1}^6 (FATURA_k - PGTO_k), \quad (1.1)$$

¹ A Recomendação de Basiléia II, acordo de capital celebrado em 1988, é de que um atraso maior do que 90 dias seja indicativo de *default*.

Tabela 2 – Categorização da variável Idade.

Categorias	Taxa de Descumprimento Média
≤ 25	26,9%
$]25, 35]$	20,0%
$]34, 45]$	22,0%
> 45	25,6%

em que k é o índice referente ao mês da coleta de dados. Pode-se fazer a seguinte interpretação desta variável:

- $SOMA_SALDO_DEVEDOR = 0$: significa que o cliente efetuou, em todos os meses, o pagamento integral da fatura, ou não utilizou o limite disponível;
 - $SOMA_SALDO_DEVEDOR > 0$: indica que o cliente, em pelo menos um dos meses, efetuou o pagamento de um valor entre o mínimo e o total da fatura ou não efetuou o pagamento;
 - $SOMA_SALDO_DEVEDOR < 0$: indica que o cliente, em pelo menos um dos meses, efetuou o pagamento de um valor maior do que o total da fatura.
- Saldo Acumulado Médio (quantitativa contínua): média da soma acumulada das diferenças mensais das variáveis que representam o saldo da fatura e o valor efetivamente pago. O cálculo da soma acumulada de 6 meses das diferenças mensais entre as variáveis $(FATURA_k)_{k=1}^6$ e $(PGTO_k)_{k=1}^6$ é feito da seguinte maneira:

$$\begin{aligned}
 SOMA_SALDO_ACUMULADO = & \\
 & (FATURA_1 - PGTO_1) + ((FATURA_1 - PGTO_1) + (FATURA_2 - PGTO_2)) + \\
 & \dots + ((FATURA_1 - PGTO_1) + (FATURA_2 - PGTO_2) + \dots + (FATURA_5 - \\
 & PGTO_5)) + ((FATURA_1 - PGTO_1) + (FATURA_2 - PGTO_2) + \dots + \\
 & (FATURA_5 - PGTO_5) + (FATURA_6 - PGTO_6)).
 \end{aligned}$$

Genericamente, o saldo devedor acumulado médio de 6 meses é dado por:

$$SALDO_ACUM_MEDIO = \frac{SOMA_SALDO_ACUMULADO}{6}. \quad (1.2)$$

A interpretação desta variável é semelhante à da variável Soma de Saldo Devedor.

- *Limit Ratio* (quantitativa contínua): saldo devedor acumulado médio dividido pelo limite, e representa a quantia que um cliente deve em seus cartões em comparação com seu limite de crédito. Caso o cliente tenha mais de um cartão, as dívidas e os limites são somados para compor o índice. Valores baixos desta variável indicam que o limite disponível é alto e o saldo devedor é baixo. Ao contrário, um percentual alto indica que o limite disponível

é baixo e o saldo devedor é alto. Em outras palavras, esta variável revela quanto do limite está comprometido. Genericamente, tem-se que:

$$\text{LIMIT_RATIO} = \frac{\text{SALDO_ACUM_MEDIO}}{\text{LIMITE}}. \quad (1.3)$$

Esta relação é importante tanto para o tomador de crédito quanto para a instituição provedora porque mostra como está sendo gerenciado o limite de crédito disponível.

- **Atraso Máximo** (qualitativa ordinal): número máximo de meses de atraso registrado para pagamento de uma fatura, por cliente. Apesar do nome escolhido, esta variável não representa um atraso nos casos em que assume valores iguais a -2 , -1 ou 0 . Genericamente:

$$\text{ATRASSO_MAXIMO} = \max_{k=1}^6 (\text{STATUS_}k).$$

Após os tratamentos realizados, a base de dados final abarca 29.601 observações e 31 variáveis, incluindo o ID do cliente e a variável resposta. Nenhum dos atributos possui valores faltantes (*missing*).

Finalmente, o objetivo desta pesquisa é utilizar diferentes algoritmos de aprendizado de máquina para treinar classificadores em um conjunto de dados de treinamento e, posteriormnete, utilizá-los para fazer previsão em um conjunto de exemplos separados para teste. Métricas de desempenho para problemas de classificação de duas classes serão utilizadas para avaliação e comparação dos resultados produzidos pelos classificadores. Os conjuntos de dados considerados englobam as variáveis preditoras especificadas nesta seção após passarem pela etapa de seleção de variáveis, descrita em [Subseção 3.1.1](#).

ANÁLISE EXPLORATÓRIA

A análise descritiva é apontada como a fase inicial do processo de estudo dos dados coletados. É a etapa em que os dados serão sintetizados através da aplicação de fórmulas estatísticas e técnicas de visualização. A análise das características por meio de medidas de estatística descritiva e de métodos gráficos permite a compreensão acerca do comportamento dos dados e a descoberta de padrões e tendências que podem ser úteis nas etapas seguintes do processo de extração de conhecimento dos dados. Este capítulo é dedicado à redução, análise e interpretação de algumas das variáveis originais da base de dados e das variáveis que foram construídas a partir delas. As ferramentas descritivas utilizadas envolvem essencialmente vários tipos de gráficos.

2.1 Análise Univariada

Medidas de Posição Central

Conjuntos de dados podem ser resumidos apresentando-se um ou alguns valores que sejam representativos da série toda. Ao usar um único valor, obtem-se uma redução drástica dos dados, o que pode ocasionar certas fragilidades. Segundo (FACELI *et al.*, 2011), medidas de localidade ou medidas de posição definem pontos de referência nos dados e as mais usualmente empregadas para atributos numéricos são a média aritmética, mediana e percentil e, para atributos categóricos, a moda.

Média Aritmética

Considerando um atributo com n registros assim como definido no [Capítulo 1](#), a média pode ser expressa por meio de um somatório dos n elementos, como mostra a [Equação 2.1](#):

$$\text{Média}(x^j) = \bar{x}^j = \frac{1}{n} \sum_{i=1}^n x_i^j \quad (2.1)$$

A média conduz a erros de interpretação, pois é uma medida sensível à presença de *outliers*, que são valores muito maiores ou menores comparativamente aos demais. A média, também representada por μ , é uma boa estimativa do meio de um conjunto de valores apenas se os valores estão simetricamente distribuídos.

Mediana

Por ser menos sensível à *outliers*, a mediana é uma alternativa à média. Para que esta medida seja utilizada, o atributo numérico precisa ser ordenado de forma crescente. Isto feito, a mediana é calculada segundo a [Equação 2.2](#), como mostram [Faceli et al. \(2011\)](#):

$$\text{Mediana}(x^j) = \begin{cases} \frac{1}{2} (x_r + x_{r+1}), & \text{se } n \text{ for par} \\ x_{r+1}, & \text{se } n \text{ for ímpar} \end{cases} \quad (2.2)$$

Ou seja, se o número de observações de um atributo é ímpar, a mediana é exatamente igual ao valor central do conjunto ordenado. Caso contrário, a mediana é dada pela média dos valores centrais.

Quartis e Percentis

Os percentis e os quartis são medidas também muito utilizadas e, tal como a mediana, os valores precisam estar ordenados para que estas estatísticas sejam calculadas. Similarmente à mediana, que divide os dados ao meio, os quartis dividem os dados em quatro partes iguais. O primeiro quartil de uma sequência, denominado $Q1$, é o valor que deixa 25% dos demais valores abaixo dele. O terceiro quartil, denominado $Q3$, é o valor que deixa 75% dos demais valores abaixo dele. A mediana é igual ao segundo quartil, $Q2$.

Percentis são valores que dividem uma série de dados ordenados em 100 partes iguais. Para [Faceli et al. \(2011\)](#), dado que p é um valor entre 0 e 100, o p° percentil, ou Pp , é um valor x_i do conjunto de valores tais que $p\%$ dos valores observados para determinado atributo são menores do que x_i . A mediana equivale também ao 50° percentil, denominado $P50$, e o primeiro e terceiro quartis equivalem, respectivamente, ao $P25$ e $P75$.

[Bussab e Morettin \(1987\)](#) estipulou as chamadas estatísticas de ordem, medidas resistentes de posição que são formadas, dentre outras medidas, pela mediana, pelos valores extremos de

um conjunto e pelos quartis. Para os autores, uma medida de dispersão ou posição é resistente se for pouco afetada por mudanças de uma pequena proporção dos dados.

Moda

A moda de um atributo qualitativo é dada simplesmente pelo valor que se repete mais vezes, ou seja, o mais frequente. Conjuntos de dados podem ter mais de uma moda e, nestes casos, são chamados de bimodais, trimodais ou multimodais. Um conjunto de dados também pode não apresentar um valor mais frequente, e neste caso é chamado de amodal.

Uma forma simples de visualizar a distribuição dos dados é utilizar técnicas de visualização, como *boxplots*, histogramas e *scatter plots*, que serão apresentados na [Seção 2.1](#).

Medidas de Espalhamento

A redução de um atributo numérico por meio de uma medida de localidade oculta informações sobre a variabilidade do conjunto de valores. As medidas de espalhamento ou dispersão, por outro lado, medem a dispersão ou espalhamento de um conjunto de valores, ainda que sejam expressas por um número. Por meio destas medidas pode-se observar se os valores de um atributo numérico estão amplamente espalhados ou concentrados em torno de um valor. As medidas de dispersão mais comumente utilizadas são a amplitude, a variância e o desvio-padrão.

Amplitude

A amplitude mensura o espalhamento máximo, ou distância máxima, entre os valores de um atributo. Assim, para um conjunto de valores x_i^j , com $i = 1, \dots, n$, a amplitude é dada pela [Equação 2.3](#):

$$\text{Amplitude}(x^j) = \max_i (x_i^j) - \min_i (x_i^j) \quad (2.3)$$

A amplitude também não leva em consideração todos os valores do conjunto de dados, como seria preferível. Segundo [Faceli et al. \(2011\)](#), dado que somente os valores extremos são considerados no cálculo, se a maioria dos valores for próxima de um ponto, a amplitude não será uma boa medida de espalhamento.

Intervalo Interquartil

O intervalo interquartil, assim como o nome sugere, representa a diferença entre o terceiro e primeiro quartis e indica a extensão do intervalo em que estão concentrados os dados do meio

de uma distribuição de dados. Bem como a amplitude, esta medida não leva em consideração todos os dados do conjunto. Sua expressão é dada pela [Equação 2.4](#):

$$IQ(x^j) = P_{75\%} - P_{25\%} \quad (2.4)$$

Desvios e Variância

Medidas de dispersão amplamente utilizadas são a variância e o desvio-padrão, os quais mensuram a concentração dos dados em torno da média. Estas medidas consideram toda a informação disponível. Num geral, o propósito é avaliar o quanto cada uma das observações de um determinado atributo x^j se desvia da média $\bar{x}^j$ do conjunto. A variância é dada pela [Equação 2.5](#):

$$\text{Variância}(x^j) = \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i^j - \bar{x}^j)^2 \quad (2.5)$$

[Bussab e Morettin \(1987\)](#) afirmam que a variância pode causar problemas de interpretação por ser uma medida que expressa um desvio quadrático médio. O fato de o cálculo da variância envolver uma potência ao quadrado aumenta a divergência entre os valores e dificulta a interpretação desta medida. Por isto, costuma-se usar o desvio-padrão, que é a raiz quadrada da variância e, portanto, uma medida expressa na mesma unidade dos valores do conjunto de dados. O desvio-padrão é então definido pela equação [Equação 2.6](#):

$$\text{Desvio}(x^j) = (\sigma^j)^2 = \sqrt[2]{\text{var}(x^j)} \quad (2.6)$$

Os valores de variância e desvio-padrão, no entanto, também são distorcidos pela presença de *outliers*, já que para compor estas medidas são calculadas as diferenças entre os valores de cada observação e a média do conjunto. Segundo a definição de [Bussab e Morettin \(1987\)](#), portanto, variância e desvio não podem ser consideradas estatísticas resistentes. Além disso, estas duas medidas estatísticas não fornecem informações acerca da assimetria da distribuição dos dados.

Coeficiente de Variação

A variância e o desvio-padrão só podem ser utilizados para comparar a variabilidade de dois ou mais conjuntos de dados quando estes apresentarem a mesma média, mesmo número de observações e estiverem expressos nas mesmas unidades. Para comparar qualquer conjunto de dados quanto à sua variabilidade deve-se utilizar uma medida de dispersão denominada coeficiente de variação, que expressa a variabilidade dos dados em relação à sua média, ou seja, é a razão do desvio-padrão pela média. O coeficiente de variação é uma medida adimensional, ou seja, não depende da unidade de medida. Essa característica permite então usá-lo para comparar

a variabilidade de conjuntos de dados medidos em unidades diferentes, o que seria impossível usando o desvio-padrão. Sua expressão é dada pela [Equação 2.7](#):

$$CV(x^j) = \frac{\sigma}{\mu} \quad (2.7)$$

O coeficiente de variação mostra qual é a contribuição do desvio-padrão σ na média μ . Quanto menor o coeficiente de variação de um conjunto de dados, menor é a sua variabilidade. Diferentes classificações para o coeficiente de variação podem ser encontradas na literatura.

Métodos Gráficos

Gráfico em Barras

Gráficos em barras são utilizados para representar as distribuições de variáveis categóricas nominais ou ordinais e as frequências absolutas ou relativas podem ser utilizadas para exprimir a representatividade de indivíduos em cada categoria.

A [Figura 1](#) exhibe as proporções de clientes adimplentes (não-*default*) e inadimplentes (*default*) na base total e de acordo com as categorias das variáveis Sexo, Escolaridade e Estado Civil. Por meio dos gráficos, nota-se que a parcela de clientes inadimplentes é significativamente inferior à de clientes adimplentes, cerca de 3,5 vezes menor. Ainda, os resultados da [Figura 1](#) indicam que a taxa de *default* para o sexo feminino é inferior à taxa registrada para o sexo masculino e que a taxa de *default* dos clientes que declararam ser casados é superior à taxa dos clientes solteiros. A taxa de *default* dos clientes que possuem um nível de escolaridade mais alto também é inferior à a taxa das outras duas categorias exibidas.

A [Figura 2](#), por sua vez, apresenta a distribuição da variável Atraso Máximo e indica que o pior cenário para aproximadamente 49% dos clientes, considerando o histórico de seis meses de informações, é o acionamento do crédito rotativo; 24% dos clientes atrasaram por no máximo dois meses o pagamento de suas faturas; 5,7% dos clientes atrasaram por no máximo um mês; menos de 1,5% dos clientes atrasaram o pagamento por quatro meses ou mais. Por outro lado, mais de 17% dos clientes não utilizaram o cartão em nenhum dos meses ou efetuaram o pagamento total de suas faturas. A observação dos resultados condicionados ao atributo alvo, como mostra a [Figura 3](#), revela que a expressividade de clientes inadimplentes é maior do que a de clientes adimplentes nos casos em que a variável Atraso Máximo assume valores maiores ou iguais a 3.

Histogramas

O histograma, também conhecido como distribuição de frequências, é um gráfico em barras utilizado para representar a distribuição de uma variável quantitativa contínua, e possibilita uma compreensão melhor do padrão geral assumido pelos dados.

Figura 1 – Distribuições da variável resposta e das variáveis categóricas sexo, escolaridade e estado civil dentre as classes da variável resposta.

Fonte: Elaborado pelo autor.

O histograma da variável Limite é retratado pela [Figura 4](#) e sugere que as frequências mais altas, ou maiores concentrações de clientes, estão associadas a limites mais baixos e que, no geral, à medida que aumentam os valores de limite de crédito, diminuem as frequências absolutas. O gráfico ao lado direito da [Figura 4](#) exibe as frequências em escala logarítmica, a qual reduz a representação a uma escala mais fácil de ser visualizada, e mostra que as frequências nos três primeiros intervalos do histograma, ou seja, de limites até 300.000 NT *dollar*, podem ser até 100 vezes maiores do que as frequências atreladas a limites superiores a 600.000 NT *dollar*.

Figura 2 – Distribuição da variável atraso máximo.

Fonte: Elaborado pelo autor.

O histograma da variável Idade, representado pela figura [Figura 5](#), indica que as maiores frequências estão compreendidas no intervalo de 20 a 30 anos e, a partir deste ponto, começam a decrescer até atingir patamares mínimos - a partir da idade de 60. Na escala linear a impressão que se tem é que a maior diferença entre duas frequências absolutas consecutivas ocorre entre as faixa de 20-22 e 22-24 anos, mas por meio da escala logarítmica é possível observar que relativamente também há uma diferença considerável, e ainda maior, entre as frequências das faixas de idade de 68-70 e 70-72 anos. Novamente a escala logarítmica é útil, pois a distribuição de frequências a partir da idade de 60 é de difícil visualização.

A distribuição de frequências da variável Saldo Acumulado Médio, apresentada na [Figura 6](#), mostra que a maior concentração de clientes ocorre no intervalo de -500.000 a 500.000 NT *dollar* e, tomando por base a escala linear, estas frequências parecem estar associadas a valores positivos da variável, com uma baixa frequência caindo na parte negativa do intervalo. A escala logarítmica, no entanto, aponta que a concentração de clientes na faixa de -100.000 a 0.000 é ainda ligeiramente superior à concentração de clientes na faixa de 100.000 a 200.000 NT *dollar* e que a frequência mais alta, de mais de doze mil clientes em termos absolutos, está associada ao intervalo 0.000 a 100.000. A escala logarítmica também mostra que as frequências

Figura 3 – Distribuição da variável atraso máximo condicionada à variável resposta.

Fonte: Elaborado pelo autor.

atreladas a saldos superiores a 2.000.000 NT *dollar* são até mil vezes menores do que as maiores frequências registradas. De maneira geral é possível observar que as frequências diminuem com o aumento do saldo acumulado médio. O histograma da variável Saldo Devedor Total é semelhante ao da variável Saldo Acumulado Médio.

Box-Plot

O *boxplot* é uma maneira simples de visualizar os dados, pois exibe simultaneamente algumas características da distribuição, como a dispersão dos dados, mediana e valores atípicos (*outliers*). A construção deste gráfico é baseada nos quartis da distribuição e na diferença interquartílica, definida em [Seção 2.1](#).

O *boxplot* à esquerda da [Figura 7](#) aponta que a variável Soma de Saldo Devedor apresenta dois *outliers* negativos e uma série de *outliers* positivos e, por conta destas observações discrepantes, a variabilidade total dos dados é grande - menor para os clientes *default*. As distâncias interquartílicas dos dois *boxplots* à esquerda sinalizam que a variabilidade da parte central dos dados é pequena e que a diferença entre a mediana e o primeiro quartil é menor do que a diferença entre o terceiro quartil e a mediana, indicação de assimetria dos dados. Além disso, as medianas

Figura 4 – Histograma da variável Limite em escala linear (ao lado esquerdo) e em escala logarítmica (ao lado direito).

Fonte: Elaborado pelo autor.

Figura 5 – Histograma da variável Idade em escala linear (ao lado esquerdo) e em escala logarítmica (ao lado direito).

Fonte: Elaborado pelo autor.

das distribuições *default* e não *default* são muito semelhantes.

A Figura 8 mostra que a variável Limite também possui uma série de valores atípicos, sendo o maior deles igual a 1.000.000 NT *dollar*. A mediana da distribuição dos clientes

Figura 6 – Histograma da variável Saldo Acumulado Médio em escala linear (ao lado esquerdo) e em escala logarítmica (ao lado direito).

Fonte: Elaborado pelo autor.

Figura 7 – *Boxplot* das variáveis Soma de Saldo Devedor e Saldo Acumulado Médio, por classe da variável resposta.

Fonte: Elaborado pelo autor.

não *default* está aproximadamente à mesma distância dos primeiro e terceiro quartis, mas as diferenças entre o terceiro quartil e o limite superior e entre o primeiro quartil e limite inferior indicam que os dados da distribuição não *default* não são simétricos. A mediana dos clientes

default está mais próxima do primeiro quartil do que do terceiro, indicando também assimetria. O gráfico ao lado direito da Figura 8 indica que a distribuição dos clientes não *default* da variável *Limit Ratio* possui mais valores discrepantes e maiores comparativamente à distribuição dos clientes *default* desta variável. Isso significa que um uso exarcebado do limite do cartão ocorre mais para clientes que não foram marcados em *default* no mês avaliado. Novamente, as distâncias entre a mediana e cada um dos os quartis não são simétricas, sendo ainda mais assimétrica para os dados dos clientes não *default*.

Figura 8 – Boxplot das variáveis Limite e *Limit Ratio*, por classe da variável resposta.

Fonte: Elaborado pelo autor.

A Figura 9 exibe os *boxplots* das variáveis Saldo Acumulado Médio e Limite de acordo com as categorias dos atributos Sexo e Idade. O gráfico ao canto superior esquerdo indica que o maior valor discrepante está associado à uma idade > 45 anos e que a variabilidade dos dados é menor para idades > 35 e ≤ 45 anos, indicando que esta categoria de clientes recebe limites menores. As distribuições dos dados centrais das categorias ≤ 25 e ≤ 35 parecem ser aproximadamente simétricas, indicando que estas duas categorias apresentam distribuições de limites muito semelhantes. O gráfico ao canto superior direito sugere que as distribuições da variável limite dentre os sexos são muito semelhantes, sendo o maior valor atípico pertencente ao sexo feminino. O gráfico ao canto inferior esquerdo mostra que a variabilidade dos dados é pequena para as categorias da variável Idade e apenas as categorias ≤ 25 e ≤ 35 possuem *outliers* negativos. No entanto, todas as categorias apresentam diversos *outliers* positivos, apontando que todas as faixas de idade comportam clientes com dívidas altas. Além disso, todas as distribuições são assimétricas. Assim como para a variável Limite, as distribuições dentre os sexos são semelhantes, assimétricas e apenas o sexo feminino apresenta valores atípicos negativos.

Figura 9 – *Boxplot* das variáveis Limite e Saldo Acumulado Médio por categorias das variáveis Sexo e Idade.

Fonte: Elaborado pelo autor.

2.2 Análise Multivariada

Medidas de Espalhamento

Covariância e Correlação

De acordo com [Faceli *et al.* \(2011\)](#), o espalhamento de atributos quantitativos de um conjunto de dados multivariado pode ser avaliado por uma matriz de covariância em que cada elemento da matriz representa a covariância entre atributos x^i e x^j . A covariância mede o grau com que dois elementos variam juntos e é dada pela equação [Equação 2.8](#):

$$\text{Covariância}(x^i, x^j) = \frac{1}{n-1} \sum_{k=1}^n (x_k^i - \bar{x}^i) (x_k^j - \bar{x}^j), \quad (2.8)$$

em que $\bar{x}^i$ e $\bar{x}^j$ representam as médias aritméticas dos atributos i e j , respectivamente, e x_k^i e x_k^j representam os valores do i -ésimo e do j -ésimo atributo para a k -ésima observação. É importante frisar que a covariância entre dois atributos iguais é igual a variância deste atributo, ou seja,

$\text{cov}(x^i, x^j) = \text{var}(x^i)$ quando $i = j$. Assim, a diagonal da matriz de covariância é igual a variância dos atributos.

Ainda segundo os autores, a medida de covariância é afetada pela magnitude dos dados. Por isso, duas variáveis com magnitudes maiores podem apresentar covariância maior do que duas variáveis de magnitude menor, mas que são mais semelhantes entre si. Não é indicado, portanto, avaliar o relacionamento entre duas variáveis através da covariância entre elas. A medida de correlação é uma alternativa a este problema por retirar a influência da magnitude dos dados, apresentando com mais clareza a força da relação linear entre dois atributos. Assim como a matriz de covariância, cada elemento de uma matriz de correlação representa a medida de correlação entre os atributos x^i e x^j . Esta medida, também conhecida como correlação linear de Pearson, pode ser definida como:

$$\text{Correlação}(x^i, x^j) = \frac{\text{Covariância}(x^i, x^j)}{\sigma^i \sigma^j}, \quad (2.9)$$

em que σ_i e σ_j são os desvios-padrão dos valores dos atributos x^i e x^j . No caso da matriz de correlação, os elementos da diagonal principal têm valor igual a 1. Os demais elementos podem assumir valores no intervalo $[-1, 1]$, como será abordado mais adiante. De um modo geral, a quantificação do grau de dependência entre duas variáveis é feita pelos chamados coeficientes de associação ou correlação. Estas são medidas que descrevem em um único número a dependência entre as duas variáveis (BUSSAB; MORETTIN, 1987).

Correlação de Pearson

O tipo mais simples de associação entre variáveis quantitativas contínuas é a linear, medida que avalia o quão próximo de uma reta é a nuvem de pontos formada pelos pares ordenados de dois atributos. Um relacionamento linear ocorre quando a mudança em uma variável é associada à uma mudança direta ou indiretamente proporcional na outra variável. Um dos coeficientes mais populares para quantificação do relacionamento linear entre duas variáveis é o coeficiente de Pearson, cujo método padrão de cálculo foi descrito pela primeira vez por Karl Pearson. Segundo Chok (2010), a Equação 2.10 define a correlação de Pearson, denotada pela letra r :

$$r(x^i, x^j) = \frac{\sum_{k=1}^n \left((x_k^i - \bar{x}^i) (x_k^j - \bar{x}^j) \right)}{\sqrt{\sum_{k=1}^n (x_k^i - \bar{x}^i)^2 \sum_{k=1}^n (x_k^j - \bar{x}^j)^2}}, \quad (2.10)$$

em que $\bar{x}^i$ e $\bar{x}^j$ são as médias das variáveis x^i e x^j , respectivamente.

O coeficiente de correlação acima definido pode assumir qualquer valor no intervalo $[-1, 1]$. Valores iguais a 0 indicam que não há associação linear entre as duas variáveis e valores maiores do que zero indicam que há uma correlação linear positiva, ou seja, as variáveis tendem

a aumentar ou diminuir conjuntamente. Da mesma forma, valores menores do que zero indicam uma associação linear negativa, ou seja, à medida que o valor de uma variável aumenta, o valor da outra variável diminui. O sinal do coeficiente indica a direção da relação entre o par de atributos, enquanto que o valor sugere a intensidade da associação existente. Neste sentido, quanto mais forte for a associação entre as variáveis, mais próximo de 1 ou -1 será o coeficiente de Pearson. Ao contrário, quanto mais próximo de 0 for o valor do coeficiente, maior será a variação em torno da reta que melhor se ajusta aos dados. As diferentes situações relatadas são exibidas na Figura 10 e a Tabela 3 apresenta diretrizes para interpretação do coeficiente:

Tabela 3 – Interpretação do relacionamento linear de Pearson.

Valor do Coeficiente	Interpretação
$ r < 0,2$	Relacionamento Fraco
$0,2 \leq r < 0,8$	Relacionamento Moderado
$ r \geq 0,8$	Relacionamento Forte

Figura 10 – Possíveis associações lineares entre duas variáveis quantitativas e os respectivos valores do coeficiente de Pearson.

Fonte: <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>.

Vale destacar que a linearidade não é uma premissa da correlação de Pearson. No entanto, havendo indícios de que a associação entre duas variáveis não é linear, o cálculo deste coeficiente não é indicado. Gráficos de dispersão podem ser utilizados para visualização do comportamento entre duas variáveis antes que uma determinada medida de associação seja apurada. Nos casos em que o cálculo é conduzido e o coeficiente apresenta um valor baixo, indicando associação

linear fraca ou inexistência de associação linear, não significa que não há correlação entre as variáveis. A associação também pode ser do tipo não-linear e, neste caso, outros coeficientes são utilizados para quantificá-la.

Correlação de Spearman

O coeficiente de correlação de Spearman também mensura a força e a direção da associação entre duas variáveis contínuas ou ordinais. Segundo Hauke e Kossowski (2011), o coeficiente de Spearman avalia o quão bem uma função monotônica arbitrária pode descrever a relação entre duas variáveis. Em uma relação monotônica, as variáveis tendem a mudar juntas mas não necessariamente a uma taxa constante, como ocorre em um relacionamento linear. Um relacionamento monotônico, portanto, é menos restritivo do que um relacionamento linear. O gráfico ao meio da Figura 11, por exemplo, mostra um relacionamento monotônico, mas não linear. A motonicidade não é uma premissa da correlação de Spearman, assim como a linearidade não é uma premissa da correlação de Pearson. Um gráfico de dispersão novamente pode ser útil, mas a constatação gráfica de um relacionamento monotônico nem sempre é possível.

Figura 11 – Exemplos de relacionamentos monotônicos e não-monotônicos entre duas variáveis.

Fonte: <https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php>

Em princípio, o coeficiente de Spearman é simplesmente um caso especial do coeficiente de Pearson no qual os dados são classificados de acordo com sua ordem antes do cálculo. A expressão matemática do coeficiente de Spearman, como sugere Berry, Johnston e Jr. (2014), é dada pela equação Equação 2.11:

$$r_s(x^i, x^j) = 1 - \frac{6 \cdot \sum_{k=1}^n d_k^2}{n^3 - n} \quad (2.11)$$

em que d_k é a diferença entre os dois postos (*ranks*) para a k -ésima observação, ou seja, $d_k = \text{rank}(x_k^i) - \text{rank}(x_k^j)$, com $i, j = 1, \dots, d$ e $k = 1, \dots, n$.

O coeficiente de Spearman também varia no intervalo $[-1, 1]$ e o valor absoluto de r_s descreve a força do relacionamento monotônico. Segundo Chok (2010), quanto mais próximo de 0 for o valor absoluto de r_s , mais fraco é o relacionamento monotônico entre duas variáveis e quanto mais próximo de -1 ou 1 , mais forte é o relacionamento monotônico. Valores iguais a -1 ou 1 indicam uma perfeita associação entre as variáveis classificadas segundo a ordem de seus valores. O coeficiente de Spearman também pode ser igual a 0 para variáveis que são relacionadas de uma maneira não-monotônica.

Métodos Gráficos

Matriz de Correlações

A Figura 12 exibe os resultados da correlação linear de Pearson das variáveis quantitativas contínuas. Os quadrados em tom de azul mais escuro apresentam fortes correlações positivas, indicando que o relacionamento entre os pares de variáveis se aproxima de uma reta e é crescente. Os quadrados em tom de azul claro, pelo contrário, apontam que não há indícios de relacionamento linear. A Figura 13, por sua vez, exibe os resultados da correlação de Spearman das variáveis qualitativas ordinais. Neste caso, a maior parte dos pares de variáveis que representam o *status* do pagamento apresentam forte relacionamento monotônico crescente.

Gráfico de Dispersão

Um procedimento bastante utilizado para avaliar a associação entre duas variáveis quantitativas é o gráfico de dispersão, ou *scatterplot*, que representa os pares de valores dos atributos no sistema cartesiano. A percepção acerca do relacionamento linear entre as variáveis Saldo Devedor Total e Saldo Devedor Acumulado Médio, exibido na Figura 14, é imediata e decorrente da maneira como foram construídas as variáveis. Por isto, serão muito semelhantes os relacionamentos individuais entre as variáveis Saldo Devedor Total e Saldo Devedor Acumulado Médio e qualquer outra variável explicativa. Não é possível, no entanto, identificar graficamente o tipo de relacionamento entre os outros pares de variáveis. Nestes casos, é importante proceder com o cálculo dos coeficientes de correlação.

Figura 12 – Matriz das correlações de Pearson das variáveis quantitativas contínuas.

Fonte: Elaborado pelo autor.

Figura 13 – Matriz das correlações de Spearman das variáveis qualitativas ordinais.

Fonte: Elaborado pelo autor.

Figura 14 – Gráfico de dispersão das variáveis quantitativas contínuas.

Fonte: Elaborado pelo autor.

REDUÇÃO DE DIMENSIONALIDADE

Com o desenvolvimento acelerado da tecnologia, uma quantidade massiva de informações é constantemente gerada. Essas informações traduzem-se em conjuntos com inúmeras variáveis, as quais estão relacionadas à dimensão dos dados. Altas dimensões representam um desafio para os métodos de aprendizado de máquinas existentes, pois podem causar *overfitting*, resultando na degeneração de *performance* dos modelos. Para abordar o problema da maldição da dimensionalidade, como é conhecido, diversas técnicas tem sido estudadas. Este capítulo apresenta as principais técnicas de redução de dimensionalidade e aborda com mais detalhes algumas abordagens de seleção de variáveis baseadas em filtros.

3.1 Técnicas de Redução de Dimensionalidade

Segundo [Faceli et al. \(2011\)](#), as técnicas de redução de dimensionalidade podem ser divididas em:

- **Agregação:** estas abordagens mapeiam o espaço original de variáveis para um novo espaço com dimensões menores através de combinações lineares ou não lineares das variáveis do espaço original. Como estas técnicas trabalham com novos espaços de variáveis, a realização de análises acerca dos atributos gerados não é trivial. A dificuldade de se estabelecer uma ligação entre as variáveis do espaço original e as variáveis do novo espaço é um dos motivos que explicam as restrições quanto à utilização destes métodos. Podem ser citados como exemplos de técnicas de agregação a Análise de Componentes Principais (PCA), Análise Discriminante Linear (LDA) e Análise de Correlação Canônica;
- **Seleção de variáveis:** estas abordagens visam selecionar um subconjunto de variáveis relevantes do conjunto original sem que nenhuma transformação seja realizada, o que mantém o significado físico e propriedades dos atributos originais. Dada a maior facilidade de compreensão, interpretação e também de implementação, o uso destas técnicas é mais

difundido. Técnicas de seleção incluem o coeficiente de Fischer Escore, *Information Gain*, coeficientes de correlação, etc.

Os dois tipos de abordagens apresentados acima são capazes de melhorar o desempenho dos modelos e diminuir a complexidade computacional. Para [Faceli et al. \(2011\)](#), em várias aplicações é importante preservar os valores dos atributos para que os resultados obtidos possam ser interpretados. Por isto, em algumas áreas, como a de finanças, é mais frequente a redução do número de atributos pelo uso de técnicas de seleção.

3.1.1 Seleção de Variáveis

De acordo com [Aggarwal \(2014\)](#), a depender se o conjunto de treinamento contém rótulos ou não, os algoritmos de seleção de variáveis podem ser categorizados em: supervisionados, semi-supervisionados e não-supervisionados. Algoritmos de seleção de variáveis supervisionados podem ainda ser amplamente categorizados em três tipos: modelos baseados em filtros, modelos embutidos e modelos *wrapper*. [Gu, Li e Han \(2011\)](#) e [Zhao et al. \(2010\)](#) também fazem esta distinção. Na abordagem embutida, a seleção do subconjunto de variáveis é integrada no próprio algoritmo de aprendizado de máquina, ou seja, a seleção de atributos é executada juntamente com o modelo de aprendizado (por exemplo: árvores de decisão). A abordagem baseada em *wrapper* usa a precisão preditiva de um algoritmo de aprendizagem pré-determinado para definir a qualidade das variáveis selecionadas. Estes métodos apresentam bom desempenho, mas são custosos em termos de tempo e complexidade computacional. Por último, na abordagem baseada em filtro um subconjunto de variáveis é selecionado independentemente do algoritmo de aprendizado de máquina que utilizará este subconjunto. Esta é uma característica que, segundo [Faceli et al. \(2011\)](#), configura-se como uma desvantagem, pois o viés de um não atinge o viés do outro, o que pode levar a construção de classificadores com desempenho inferior ao desejado. [Aggarwal \(2014\)](#) destaca que os modelos baseados em filtros dependem das características gerais dos dados, como distâncias, informações, correlações. Estes modelos costumam apresentar rápido processamento computacional, podendo ser utilizados em conjuntos que apresentam grande quantidade de dados.

Segundo [Zhao et al. \(2010\)](#), em problemas de classificação, devido à disponibilidade de informações de rótulos, a relevância das variáveis é avaliada como a capacidade de distinguir as diferentes classes do atributo alvo. [Aggarwal \(2014\)](#) coloca que na maioria dos problemas reais pouco conhecimento se têm a respeito da relevância das variáveis. Portanto, na tentativa de melhor representar o problema, muitas variáveis são introduzidas, e muitas destas variáveis são redundantes ou irrelevantes. Uma variável irrelevante não está diretamente associada à variável resposta, mas afeta o processo de aprendizado. Uma variável redundante, por outro lado, nada acrescenta à variável resposta. A presença de variáveis redundantes ou irrelevantes afeta o desempenho dos métodos de classificação. A redução de atributos redundantes ou irrelevantes pode otimizar o tempo de processamento e produzir um classificador melhor.

3.1.1.1 Métodos de Seleção de Variáveis Baseados em Filtros

Um típico modelo de seleção de variáveis baseado em filtro, de acordo com [Aggarwal \(2014\)](#), consiste essencialmente em duas etapas: na primeira as variáveis são ordenadas segundo algum critério e, na segunda, as variáveis avaliadas na primeira etapa e mantidas para análise são selecionadas para induzir os métodos de classificação. Uma variável desconsiderada na primeira etapa pode ser útil se combinada com outras variáveis. Informações como estas não são capturadas com métodos univariados. No esquema univariado cada variável é ordenada independentemente do espaço de variáveis, enquanto que no esquema multivariado as variáveis são selecionadas de maneira conjunta. Um esquema multivariado é capaz de evitar a redundância entre as variáveis, enquanto que o univariado não. Os métodos apresentados a seguir envolvem basicamente o cálculo de correlações. As variáveis são avaliadas, portanto, de forma univariada.

Teste Qui-Quadrado de Pearson para Dados Nominais e Medidas de Associação Baseadas na Estatística Qui-Quadrado

Uma das medidas que quantifica a dependência entre duas variáveis qualitativas é chamada de qui-quadrado de Pearson, estatística não-paramétrica que foi projetada para analisar as diferenças de grupos quando as variáveis são mensuradas em um nível nominal. O teste qui-quadrado de independência é um teste de hipótese que objetiva avaliar a associação entre duas variáveis qualitativas nominais através da comparação das frequências observadas e esperadas para um evento. Testes de hipótese normalmente são aplicados na busca por evidências estatísticas de que duas variáveis possuem determinado grau de associação. Busca-se o confronto entre duas hipóteses, a nula e a alternativa, que neste contexto correspondem à existência ou inexistência de associação. Pode-se dizer que dois grupos se comportam de forma semelhante se as diferenças entre as frequências esperada e observada em cada categoria forem muito pequenas. Suponha que se queira avaliar, por exemplo, a dependência entre o atributo alvo e a variável indicadora do sexo do cliente. Para isto, considera-se os dados da [Tabela 4](#):

Tabela 4 – Distribuição conjunta dos valores absolutos e relativos dos clientes de acordo com o sexo e as categorias da variável resposta.

Classes da Variável Resposta	Sexo		Total
	Feminino	Masculino	
0	14.111 (79%)	8.885 (76%)	22.996 (78%)
1	3.744 (21%)	2.861 (24%)	6.605 (22%)
Total	17.855 (100%)	11.746 (100%)	29.601 (100%)

É possível observar que, independentemente do sexo, 78% dos clientes efetuaram o pagamento de suas faturas no mês avaliado e 22% dos clientes não efetuaram. Não havendo dependência entre as variáveis, espera-se estas mesmas proporções para cada sexo. A observação dos resultados indica que as proporções dos sexos feminino (79% e 21%) e masculino (76% e 24%) são

próximas, mas não são exatamente iguais às proporções esperadas, que podem ser calculadas da seguinte maneira:

$$E_{l,c} = \frac{(\text{Total de Linhas}) * (\text{Total de Colunas})}{\text{Total da Amostra}}. \quad (3.1)$$

A [Tabela 5](#) apresenta os valores absolutos correspondentes às frequências esperadas em cada célula, calculados segundo a equação [Equação 3.1](#). Comparando as duas tabelas, parece haver diferença, ainda que pequena, entre os valores observados e os esperados, mas para quantificar esta diferença é necessário estabelecer uma medida.

Tabela 5 – Valores esperados na tabela [Tabela 4](#) supondo independência entre as variáveis.

Classes da Variável Resposta	Sexo		Total
	Feminino	Masculino	
0	13.871 (78%)	9.125 (78%)	22.996 (78%)
1	3.984 (22%)	2.621 (22%)	6.605 (22%)
Total	17.855 (100%)	11.746 (100%)	29.601 (100%)

A estatística do teste qui-quadrado de Pearson é conhecida como uma medida de afastamento global e mensura justamente esta relação entre os valores esperados e observados da distribuição conjunta de duas variáveis. Sua expressão é dada por:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad (3.2)$$

em que O_{ij} e E_{ij} representam, respectivamente, as frequências absolutas observadas e esperadas da i -ésima linha e da j -ésima coluna. Quando as frequências esperadas são muito próximas às frequências observadas, o valor de χ^2 é pequeno. Quando as divergências são grandes, da mesma forma passa a ser a estatística qui-quadrado, indicando dependência entre as variáveis. Segundo [Bussab e Morettin \(1987\)](#), quanto maior for o valor de χ^2 , maior será a associação entre as variáveis. A qui-quadrado representada pela [Equação 3.2](#) é a estatística de teste e a hipótese nula de que as variáveis são independentes é confrontada com a hipótese alternativa de que as variáveis não são independentes.

O teste qui-quadrado de independência não é uma medida de força da associação entre as variáveis. [Agresti \(2003\)](#) afirma que um p-valor pequeno indica evidência de associação entre as variáveis, mas provê poucas informações acerca da natureza ou força da associação. [Bussab e Morettin \(1987\)](#) destacam que, dada a dificuldade de julgar se a associação entre as variáveis é alta ou não baseando-se na estatística qui-quadrado, Pearson propôs o chamado coeficiente de contingência C , definido pela [Equação 3.4](#):

$$C = \left(\frac{\chi^2}{\chi^2 + n} \right)^{\frac{1}{2}}, \quad (3.3)$$

Os autores colocam que, ainda que exista uma associação perfeita entre as variáveis, o coeficiente de contingência acima definido pode não ser igual a 1. Por isto, a seguinte alteração é sugerida, em que t é igual ao mínimo entre o número de linhas e colunas da tabela:

$$C^* = \left(\frac{t\chi^2}{(t-1)(\chi^2 + n)} \right)^{\frac{1}{2}}, \quad (3.4)$$

Este coeficiente é limitado inferiormente por 0 e superiormente por 1, assumindo valor igual a 0 quando não há associação entre as variáveis.

Outros coeficientes também podem ser encontrados na literatura. Para o teste qui-quadrado, [Liebetrau \(1983\)](#) apresenta o coeficiente de Cramer, proposto por Cramer em 1946 e que pode ser calculado por meio da [Equação 3.5](#). Em caso de associação perfeita, $v = 1$, enquanto que $v = 0$ em caso de independência.

$$V = \left(\frac{\chi^2}{n(t-1)} \right)^{\frac{1}{2}}, \quad (3.5)$$

A [Tabela 6](#) exibe os valores da estatística qui-quadrado e dos coeficientes de contingência de Pearson e Cramer para os pares de variáveis formados por cada atributo qualitativo nominal e o atributo alvo. Os resultados evidenciam que a estatística χ^2 é maior quando a variável resposta e a variável indicadora do sexo são analisadas conjuntamente. Os coeficientes de contingência próximos de zero, no entanto, mostram que a associação entre as variáveis nominais e a variável resposta é fraca. Segundo [Khamis \(2008\)](#), valores próximos de zero indicam pouca associação, e valores próximos de 1 indicam previsibilidade quase perfeita. Com base nisto, estas duas variáveis não serão mantidas para análise.

Tabela 6 – Estatística qui-quadrado de Pearson e coeficiente de correlação entre as variáveis qualitativas e a variável resposta.

Variáveis Explicativas	Variável Resposta			
	Estatística Qui-Quadrado	Valor-p do Teste Qui-Quadrado	Coeficiente de Contingência	Coeficiente de Cramer
Sexo	46,7	8,15e-12	0,06	0,04
Estado Civil	32,0	1,13e-07	0,05	0,03

O Teste Qui-Quadrado para Linearidade

O teste qui-quadrado de independência é importante para a análise de dados categóricos, mas não é indicado quando se pretende avaliar a associação entre uma variável categórica nominal e uma variável categórica ordinal ou duas variáveis ordinais. Se uma tabela de contingência for reorganizada, de maneira que uma ou mais linhas ou colunas forem trocadas, e a estatística qui-quadrado for recalculada, o resultado será igual ao primeiro. E isto é esperado, já que esta

estatística não leva em consideração a ordem das linhas ou colunas da tabela de contingência. Por vezes, no entanto, a ordem das linhas ou colunas faz diferença, ou seja, os eventos de uma variável categórica carregam alguma informação ordinal que precisa ser considerada no cálculo do coeficiente de associação. Uma abordagem sugerida por Agresti (2003), conhecida como "Qui-Quadrado para Linearidade" ou "*Linear-by-linear Association Model*", pode ser utilizada para testar a associação entre variáveis com categorias ordenadas. O autor ressalta que quando a variável X da linha e a variável Y da coluna são ordinais, uma tendência positiva ou negativa na associação é comum. A análise atribui escores às categorias, os quais determinam distâncias entre elas, respeitando a ordenação, e tratam a escala de medição como intervalar. A informação de correlação, estatística de teste que é sensível a tendências lineares, é então utilizada.

A qui-quadrado para linearidade verifica se duas variáveis ordinais, ou uma variável dicotômica e outra ordinal, são associadas. Na literatura consta que outros coeficientes de correlação, como o de Spearman ou Kendall, também podem ser conduzidos neste caso. A estatística para testar independência é:

$$M^2 = (N - 1)r^2, \quad (3.6)$$

em que r é o coeficiente de correlação de Pearson e N é o tamanho da amostra. À medida que n ou r^2 aumentam, M^2 também aumenta.

A hipótese nula para o teste de significância é que não há associação entre as variáveis da tabela. Um p-valor significativo não implica que a associação seja linear, meramente que buscar uma componente linear para a associação ajuda a construir poder contra a H_0 . Sob a hipótese de independência das variáveis, $r = 0$ e $M^2 = 0$ e sob a hipótese de perfeita associação, $M^2 = (N - 1)$. Ou seja, quanto maior é o coeficiente de correlação em valor absoluto, mais distante da independência os dados estão nesta dimensão linear.

As categorias originais da variável ordinal podem ser utilizadas ou pesos podem ser atribuídos a categorias consideradas mais importantes, como colocado anteriormente. Na prática, desde que números escolhidos aumentem monotonicamente, e não sejam muito extremos, o resultado não mudará muito em função da nossa escolha. Esta etapa é a que estabelece métricas para as variáveis independentes. A atribuição de pesos, vale ressaltar, é arbitrária. No *software* R a função `lbl_test` do pacote *coin* trata automaticamente as variáveis como ordenadas, com os níveis na tabela ordenados por ordem crescente. Por *default*, os níveis são igualmente espaçados, mas a opção de *scores* pode ser utilizada para especificar as distâncias entre os níveis das variáveis, ou seja, para atribuir alguma ponderação às classes.

O qui-quadrado para linearidade é também um teste de significância, não uma medida de força da associação. Os coeficientes de Pearson e Cramer, discutidos na Seção 3.1.1.1, são medidas de correlação empregadas quando se tem disponíveis variáveis nominais. Neste sentido, Khamis (2008) e Berry, Johnston e Jr. (2014) destacam que um coeficiente de correlação

denominado *rank biserial* e proposto em 1956 por Edward Cureton pode ser utilizado como medida de associação se uma das variáveis é nominal e a outra é ordinal. Segundo [Berry, Johnston e Jr. \(2014\)](#), Cureton estabeleceu que o coeficiente r_{rb} , como é designado, é claramente do tipo Kendall, e também é do tipo Spearman, fato este que tem sido investigado por pesquisadores. Isto significa que é possível estabelecer uma relação entre estes dois coeficientes e o *rank biserial*. Neste trabalho, no entanto, será exibido o resultado do teste de significância expresso pela [Equação 3.6](#) e não do coeficiente de correlação, dada a dificuldade de encontrar aplicações e explicações mais detalhadas na literatura.

A [Tabela 7](#) apresenta os resultados do teste de significância qui-quadrado para dados ordinais. Os p-valores do teste indicam que há evidências de associação entre a variável resposta e as variáveis ordinais. Desta análise, e considerando que a associação entre as variáveis não foi de fato mensurada, todas as variáveis ordinais serão selecionadas para a próxima etapa, a de modelagem.

Tabela 7 – Significância do teste qui-quadrado para linearidade aplicado às variáveis qualitativas ordinais e a variável resposta.

Variáveis Explicativas	Variável Resposta	
	Estatística do Teste Qui-Quadrado Linear	Valor-p do Teste Qui-Quadrado Linear
STATUS_ABR05	647,97	< 2,2e-16
STATUS_MAI05	780,20	< 2,2e-16
STATUS_JUN05	1386,72	< 2,2e-16
STATUS_JUL05	1639,11	< 2,2e-16
STATUS_AGO05	2064,28	< 2,2e-16
STATUS_SEP05	3152,40	< 2,2e-16
Atraso Máximo	3260,20	< 2,2e-16
Escolaridade	71,32	< 2,2e-16
Idade Categorizada	2,22	< 1,4e-01

Análise Discriminante Linear de Fisher e a Correlação Point-Biserial

A Análise Discriminante Linear de Fisher, ou LDA, é uma abordagem eficiente bastante utilizada em estatística, reconhecimento de padrões e aprendizado de máquina, cujo objetivo é determinar um subconjunto de variáveis de modo que, no espaço de dados abrangido pelas variáveis selecionadas, as distâncias entre pontos de dados em diferentes classes são tão grandes quanto possível, enquanto que as distâncias entre pontos de dados na mesma classe são tão pequenas quanto possível ([GU; LI; HAN, 2011](#)). Em outras palavras, procura-se reduzir a variação dos dados em uma mesma classe e aumentar a separação entre as classes. De maneira geral, calcula-se a diferença entre as médias de uma variável quantitativa dadas as classes da variável resposta dividida pela raiz quadrada da soma das variâncias. Um valor mais alto indica uma melhor separação entre as distribuições das classes da *target* para a variável explicativa em questão, o que significa que esta variável possui um melhor poder preditivo. O Fisher

Ratio, como também é denominado, pode ser inclusive utilizado se a *target* for contínua e a variável explicativa for categórica. Caso a variável dependente possua mais de duas classes, pode-se realizar uma análise de variância, ANOVA. O critério de avaliação usado na Análise Discriminante Linear de Fisher é o seguinte:

$$S_F(x^j) = \frac{\sum_{c=1}^l n_c (\mu_{i_c}^j - \mu^j)^2}{\sum_{c=1}^l n_c (\sigma_{i_c}^j)^2}, \quad (3.7)$$

μ^j é a média da variável x^j para todo o conjunto de dados e n_c , $\mu_{i_c}^j$ e $(\sigma_{i_c}^j)^2$ são, respectivamente, o número de observações, a média e variância da variável x^j nas classes $c = 1, \dots, l$.

O método baseado na Análise Discriminante Linear de Fisher é um algoritmo de seleção de variáveis efetivo, que tem sido amplamente utilizado em uma variedade de aplicações reais. No entanto, este algoritmo avalia as variáveis individualmente e, por isso, falha em selecionar os atributos que apresentam pontuações individuais relativamente baixas, mas uma pontuação muito alta quando são combinados com outros atributos. Da mesma maneira, o algoritmo não lida com a redundância de variáveis.

A [Tabela 8](#) exibe o resultado do Fisher *Ratio* das variáveis quantitativas contínuas. Os resultados mostram que as variáveis Limite, Limit Ratio e $(PGTO_n)_{n=1}^6$ melhor discriminam as classes da *target*.

Tabela 8 – Fisher *Ratio* das variáveis quantitativas contínuas.

Variável	Fisher Ratio
Limite	0,07556
Limit Ratio	0,03977
PGTO_SEP05	0,02088
PGTO_AGO05	0,01336
PGTO_JUN05	0,01139
PGTO_JUL05	0,01097
PGTO_MAI05	0,01059
PGTO_ABR05	0,00984
FATURA_SEP05	0,00107
Idade	0,00058
FATURA_AGO05	0,00054
FATURA_JUL05	0,00053
FATURA_JUN05	0,00026
FATURA_MAI05	0,00011
FATURA_ABR05	0,00008
Saldo Devedor Total	0,00006
Soma Saldo Devedor Acumulado	0,00002
Saldo Acumulado Médio	0,00002

O coeficiente *Fisher Ratio* é uma medida do quanto a variável preditora discrimina as classes da variável resposta, mas não é por si só uma medida de correlação, assim como são os Coeficientes de Contingência apresentados na [Seção 3.1.1.1](#). Segundo [Berry, Johnston e Jr. \(2014\)](#), a força da associação entre uma variável qualitativa nominal e uma variável quantitativa contínua pode ser mensurada por um coeficiente de correlação chamado *Point Biserial*, que é igual ao de Pearson quando este é calculado baseado em dados em que uma das variáveis é binária e a outra é contínua. [Lira e Neto \(2006\)](#) colocam que se for atribuído o valor 1 para as observações de uma das categorias da variável nominal e 0 para outra e o coeficiente do momento produto de Pearson for calculado, então o resultado será o coeficiente *Point Biserial*, que deve ser interpretado da mesma forma que r . A correlação *Point Biserial* entre a variável resposta e o atributo x^j é dada por:

$$r_{pb}(x) = \frac{\bar{x}_{i_1}^j - \bar{x}_{i_0}^j}{\sigma^j} \sqrt{\frac{n_0 n_1}{n(n-1)}}, \quad (3.8)$$

em que n_0 e n_1 representam os totais de observações do atributo x^j associados às classes $c = 0$ e $c = 1$, respectivamente, $\bar{x}_{i_0}^j$ e $\bar{x}_{i_1}^j$ são as médias do atributo x^j associadas às classes $c = 0$ e $c = 1$ da variável resposta, respectivamente, σ^j é o desvio padrão de x^j e $n = n_0 + n_1$.

Assim como para o coeficiente de Pearson, a hipótese de que não há correlação entre as variáveis também pode ser testada. Os resultados do coeficiente de correlação e do teste de hipótese são exibidos na [Tabela 9](#) e foram ordenados de forma crescente segundo o valor absoluto da medida de correlação. É possível observar que as variáveis Limite e Limit Ratio são significativamente mais correlacionadas com a variável resposta do que as outras variáveis quantitativas preditoras. Além disso, nota-se que as ordenações obtidas através da utilização da correlação *Point Biserial* e do coeficiente *Fisher Ratio* são iguais, o que confere coerência e consistência adicionais aos resultados. Como não foram encontradas na literatura referências ou orientações informais que possibilitem a interpretação dos resultados obtidos através da aplicação destas duas metodologias, neste trabalho foi determinado que serão selecionadas para a etapa de modelagem variáveis cujo valor do coeficiente *Fisher Ratio* é maior ou igual a 0,00984 ou o valor da correlação *Point Biserial* é maior ou igual a 0,0537 em módulo, pontos em que foram identificadas as maiores reduções em termos relativos. Apesar de a correlação *Point Biserial* ser derivada do coeficiente de correlação de Fisher, não foi encontrado nenhum esclarecimento quanto à paridade da interpretação dos resultados.

Determinando Medidas de Associação Apropriadas

A determinação de medidas de associação a serem utilizadas depende estritamente dos tipos de variáveis envolvidas no problema. [Khamis \(2008\)](#) resume suas recomendações da seguinte maneira:

Tabela 9 – Correlações *Point Biserial* entre a variável resposta e as variáveis quantitativas e significância do teste.

Variáveis Explicativas	Variável Resposta	
	Estatística do Teste Point Biserial	Valor-p do Teste Point Biserial
Limite	-0,1544	0,0000
Limit Ratio	0,1182	0,0000
PGTO_SEP05	-0,0739	0,0000
PGTO_AGO05	-0,0583	0,0000
PGTO_JUN05	-0,0570	0,0000
PGTO_JUL05	-0,0563	0,0000
PGTO_MAI05	-0,0561	0,0000
PGTO_ABR05	-0,0537	0,0000
FATURA_SEP05	-0,0193	0,0009
Idade	0,0144	0,0131
FATURA_AGO05	-0,0137	0,0183
FATURA_JUL05	-0,0135	0,0202
FATURA_JUN05	-0,0095	0,1031
FATURA_MAI05	-0,0062	0,2841
FATURA_ABR05	-0,0053	0,3584
Saldo Acumulado Final	0,0045	0,4382
Soma Saldo Devedor Acumulado	0,0025	0,6617
Saldo Acumulado Médio	0,0025	0,6617

Tabela 10 – Medidas de Associação recomendadas de acordo com os tipo de variáveis envolvidas na análise.

Variável Y	Variável X		
	Nominal	Ordinal	Contínua
Nominal	Coeficiente Phi	Rank Biserial	Point Biserial
Ordinal	Rank Biserial	Kendall ou Spearman	Kendall ou Spearman
Contínua	Point Biserial	Kendall ou Spearman	Pearson ou Spearman

Conforme destaca o autor, alguns cuidados devem ser tomados ao interpretar os resultados advindos de medidas de associação. Por exemplo, de que uma forte associação entre duas variáveis não implica necessariamente um relacionamento de causa e efeito.

MÉTODOS DE CLASSIFICAÇÃO

Muitos métodos estatísticos têm sido utilizados para desenvolver modelos de previsão de risco. A tarefa de previsão, não somente na área de análise de risco, está essencialmente concentrada na classificação dos clientes com relação à sua capacidade de pagamento. Como diversas metodologias podem ser empregadas com este intuito, este capítulo apresenta a revisão bibliográfica de cinco métodos de aprendizado de máquina que podem ser utilizados em problemas de classificação.

4.1 Classificador Naive Bayes

O classificador Naive Bayes é um método de aprendizado bayesiano altamente prático que classifica um objeto em uma determinada classe baseando-se na probabilidade deste objeto pertencer a esta classe. Segundo [Mitchell \(1997\)](#), em alguns domínios seu desempenho tem se mostrado comparável ao da rede neural e da aprendizagem de árvores de decisão, chegando a superá-los em alguns casos. O raciocínio bayesiano, como afirma o autor, é muito relevante para o estudo de aprendizado de máquina e fundamenta-se no pressuposto de que as quantidades de interesse são governadas por distribuições de probabilidade e que decisões ótimas podem ser tomadas pelo raciocínio sobre essas probabilidades em conjunto com os dados observados. Apesar da relevância e praticidade, [Mitchell \(1997\)](#) destaca que uma das dificuldades práticas na aplicação de métodos como este é que tipicamente requer-se conhecimento prévio de muitas probabilidades, as quais nem sempre são conhecidas. O teorema de Bayes está no cerne do aprendizado bayesiano, pois provê uma maneira de calcular essas probabilidades - as probabilidades *a posteriori*.

Mais precisamente, ainda segundo [Mitchell \(1997\)](#), o teorema de Bayes fornece uma maneira de calcular a probabilidade de uma hipótese utilizando algum conhecimento inicial sobre a probabilidade anterior desta hipótese, a probabilidade de observar vários objetos dada esta hipótese e a própria probabilidade dos dados, ou objetos. Formalmente, considerando que h é esta

hipótese, $P(h)$ denota a probabilidade de h antes mesmo que se tenha observado os dados, ou seja, não reflete qualquer conhecimento acerca da chance de ocorrência de h e, por isso, é chamada de probabilidade *a priori*; $P(D)$ denota a probabilidade de um objeto D ser observado; e $P(D|h)$ a probabilidade de se observar o objeto D dada a hipótese h . Em problemas de aprendizado de máquinas, o interesse maior está em calcular $P(h|D)$, conhecida como probabilidade *a posteriori* de h , pois reflete a confiança acerca da ocorrência de h depois de observados os dados de treinamento D . Pelo teorema de Bayes, $P(h|D)$ é calculada da seguinte maneira:

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}. \quad (4.1)$$

Frequentemente tem-se interesse em determinar a hipótese h mais provável de um espaço H de hipóteses dados objetos de treinamento D . De acordo com a definição de [Mitchell \(1997\)](#), uma hipótese cuja probabilidade é máxima é conhecida como hipótese *maximum a posteriori* (MAP), e pode ser determinada utilizando-se o teorema de Bayes e calculando-se a probabilidade *a posteriori* de cada hipótese candidata, conforme mostra a [Equação 4.2](#):

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(h|D) \\ &= \arg \max_{h \in H} \frac{P(D|h) P(h)}{P(D)} \\ &= \arg \max_{h \in H} P(D|h) P(h). \end{aligned} \quad (4.2)$$

[Faceli et al. \(2011\)](#) afirmam que o termo $P(D)$, no penúltimo passo, pode ser omitido, pois é uma constante independente de h .

De forma importante [Mitchell \(1997\)](#) coloca que, já que o teorema de Bayes fornece uma maneira para calcular probabilidades *a posteriori* de cada hipótese dados objetos de treinamento, pode-se usá-lo como base para um algoritmo de aprendizado simples que calcula a probabilidade para cada hipótese e produz a mais provável. Trazendo as definições e formulações acima para problemas de classificação, o espaço H de hipóteses pode ser um conjunto contendo possíveis valores ou categorias de uma variável resposta e D pode representar os exemplos de treinamento.

O classificador Naive Bayes aplica-se à tarefas de aprendizagem em que cada objeto x é descrito por um conjunto de atributos e a variável explicativa pode tomar qualquer valor em um conjunto finito Y . A abordagem bayesiana para classificar novas instâncias é atribuir a resposta mais provável, designada por y_{MAP} , dados atributos $x^1, x^2, \dots, x^j$. Trazendo a [Equação 4.2](#) para este contexto, tem-se o seguinte desenvolvimento:

$$\begin{aligned}
y_{MAP} &= \arg \max_{y_c \in Y} P(y_c | x^1, x^2, \dots, x^j) \\
&= \arg \max_{y_c \in Y} \frac{P(x^1, x^2, \dots, x^j | y_c) P(y_c)}{P(x^1, x^2, \dots, x^j)} \\
&= \arg \max_{y_c \in Y} P(x^1, x^2, \dots, x^j | y_c) P(y_c). \tag{4.3}
\end{aligned}$$

As probabilidades $P(y_c)$ na [Equação 4.3](#) são simplesmente a frequência de ocorrência de cada categoria da variável resposta. Por outro lado, estimar o termo $P(x^1, x^2, \dots, x^j | y_c)$ pode não ser tão simples em grandes conjuntos de dados, pois é preciso calcular as probabilidades dos possíveis valores de cada atributo condicionadas à cada possível valor da variável resposta.

Segundo [Marsland \(2014\)](#), pode-se supor, para o classificador Naive Bayes, que os valores dos atributos são condicionalmente independentes uns dos outros, dado o rótulo. Nas palavras de [Mitchell \(1997\)](#), a suposição é que, dado o valor da variável resposta de uma instância, a probabilidade de observar a conjunção $x^1, x^2, \dots, x^j$ é simplesmente o produto das probabilidades para os atributos individuais, ou seja, $P(x^1, x^2, \dots, x^j | y_c) = \prod_j P(x^j | y_c)$. Fazendo esta substituição na [Equação 4.3](#), temos o classificador Naive Bayes:

$$y_{NB} = \arg \max_{y_c \in Y} P(y_c) \prod_j P(x^j | y_c). \tag{4.4}$$

em que y_{NB} denota o valor alvo produzido pelo classificador Naive Bayes.

[Domingos e Pazzani \(1997\)](#) apud [FACELI *et al.*, 2011](#), p. 76) mostram que o classificador Naive Bayes tem um bom desempenho em uma grande variedade de domínios, incluindo muitos em que há claras dependências entre os atributos. [Kononenko \(1991\)](#) apud [FACELI *et al.*, 2011](#), p. 76) propõem que este classificador é robusto à presença de ruídos e atributos irrelevantes.

4.2 K-Nearest Neighbor

O k -vizinho mais próximo (K-NN) é um método não paramétrico baseado no conceito de distância, ou seja, na proximidade entre os dados. Apesar de simples, é amplamente utilizado e frequentemente produz bons resultados. Segundo [Faceli *et al.* \(2011\)](#), a hipótese base é que dados similares tendem a estar concentrados em uma mesma região no espaço de entrada e, da mesma forma, dados que não são similares estarão distantes entre si. O algoritmo classifica um dado elemento de acordo com as classes dos k vizinhos mais próximos pertencentes ao conjunto de dados de treinamento. Não é difícil perceber que não há esforço computacional na fase de treinamento, já que o algoritmo não aprende um modelo para os dados, apenas memoriza os objetos de treinamento para posteriormente calcular a distância de cada um deles até os objetos

do conjunto de teste. Depois de calculadas, as distâncias até cada novo objeto são ordenadas de forma crescente e seleciona-se apenas os k primeiros elementos de treinamento para realizar a classificação.

Dois pontos importantes no k -vizinhos mais próximos são a escolha (1) da função que calcula a distância entre os objetos e (2) da regra de classificação, ou seja, do valor de k . As variações que o algoritmo pode sofrer, portanto, estão associadas aos diferentes valores que o parâmetro k pode assumir e às diferentes medidas de distância que podem ser aplicadas. Como afirmam [Faceli et al. \(2011\)](#), a versão mais simples do algoritmo K-NN é o 1-vizinho mais próximo e a métrica de distância mais utilizada é a distância euclidiana, exibida pela [Equação 4.5](#):

$$D(x_{tr}, x_{te}) = \sqrt{\sum_{j=1}^d (x_{tr}^j - x_{te}^j)^2}, \quad (4.5)$$

em que $D(x_{tr}, x_{te})$ é a distância entre os objetos x_{tr} e x_{te} dos conjuntos de treinamento e teste, respectivamente, representados por vetores cujos elementos são os valores de cada um dos d atributos. No 1-NN, como também é conhecido, a classe associada ao objeto de treinamento mais próximo é utilizada para rotular o novo objeto.

Ao invés de considerar 1 vizinho mais próximo, pode-se considerar k vizinhos mais próximos, como foi colocado. Neste caso, [Faceli et al. \(2011\)](#) afirmam que as previsões dos diferentes vizinhos são agregadas de forma a classificar o ponto de teste. Em problemas de classificação, cada vizinho vota em uma classe e o objeto de teste é classificado segundo a classe mais votada - em problemas de regressão, outras estratégias podem ser consideradas. Para [James et al. \(2013\)](#), o algoritmo K-NN é um método que, dado um inteiro positivo k e uma observação x_0 do conjunto de teste, estima a probabilidade condicional da classe c como a fração dos k pontos do conjunto de treinamento mais próximos à x_0 cuja variável resposta é igual a c ([Equação 4.6](#)) e associa à observação de teste a classe cuja probabilidade é maior:

$$Pr(y = c | x = x_0) = \frac{1}{k} \sum_{i \in v_0} I(y_i = c), \quad (4.6)$$

em que v_0 o conjunto formado pelos k vizinhos mais próximos. A [Figura 15](#) exibe um exemplo simples do uso do algoritmo dos vizinhos mais próximos para $k = 7$.

O parâmetro k é definido pelo usuário e a escolha do valor mais apropriado pode ser definida através de validação cruzada. Em problemas de classificação, vale destacar, é usual utilizar valores ímpares para evitar empates. Também pode-se associar um peso à contribuição de cada vizinho, de forma que vizinhos mais próximos têm pesos maiores.

Segundo [Faceli et al. \(2011\)](#), o KNN é um algoritmo preguiçoso, porque toda a computação é adiada até a fase de classificação. A predição, no entanto, pode ser custosa, já que é necessário calcular a distância de um novo objeto até cada objeto de treinamento. Por isso, para

Figura 15 – Exemplo de classificação do método KNN com dois rótulos de classes e $k=7$.

Fonte: <<http://www.computacaointeligente.com.br/algoritmos/knn-k-vizinhos-mais-proximos/>>.

um grande conjunto de dados esse processo pode ser demorado. Por ser um método baseado em distância, como afirmam os autores, o desempenho pode ser afetado pela presença de atributos redundantes e irrelevantes. Apesar disto, o algoritmo é simples, aplicável mesmo em problemas complexos e é incremental, ou seja, quando novos exemplos de treinamento estão disponíveis, é suficiente armazená-los na memória.

4.3 Regressão Logística

Segundo Hosmer e Lemeshow (2000), métodos de regressão tornaram-se uma componente integral de qualquer análise de dados relativa à descrição do relacionamento entre uma variável resposta e uma ou mais variáveis explicativas. Os autores destacam que são frequentes os casos em que a variável resposta é discreta, assumindo dois ou mais valores e, nestas situações, o modelo de regressão logística tornou-se o método padrão de análise. O que distingue um modelo de regressão logística de um modelo de regressão linear é que a variável de resultado na regressão logística é binária ou dicotômica. Uma vez que esta diferença é levada em consideração, os métodos empregados em uma análise que utiliza regressão logística seguem os mesmos princípios usados na regressão linear.

O modelo de regressão logística, ou modelo logito, estima a probabilidade de uma variável resposta y pertencer a uma categoria específica baseado em uma ou mais variáveis preditoras. Quando dispõe-se de apenas uma variável explicativa, tem-se uma regressão logística simples e nos casos em que se tem duas ou mais variáveis explicativas, a regressão é usualmente chamada de múltipla. Quanto à variável resposta, a regressão logística pode ser classificada

em binomial, multinomial ou ordinal. O primeiro tipo lida com situações em que a variável dependente pode assumir dois valores distintos, não necessariamente numéricos, mas que normalmente são codificados como 0 ou 1; o segundo tipo faz referência a situações em que a variável dependente pode assumir três ou mais valores distintos, também não necessariamente numéricos; e, por último, o terceiro tipo envolve uma variável resposta ordinal e, neste caso, a codificação para um valor numérico deve preservar a ordem das categorias. Qualquer uma destas abordagens faz uso de uma ou mais variáveis independentes que podem ser categóricas ou contínuas, devendo cada qual receber o adequado tratamento - este ponto será discutido com mais detalhes no [Capítulo 6](#).

O modelo logito não realiza classificação, mas pode ser utilizado como um classificador escolhendo-se um ponto de corte e atribuindo uma das classes às entradas cuja probabilidade estimada é maior do que o ponto de corte escolhido e, da mesma maneira, atribuindo a outra classe às entradas cuja probabilidade estimada é menor do que o ponto de corte.

Na regressão logística binomial, a qual será considerada neste trabalho, o resultado notável, referido como sucesso, é codificado para 1 e, ao contrário, o resultado referido como fracasso é codificado para 0. Normalmente, tem-se interesse em modelar os casos de sucesso. Em outras palavras, esta metodologia modela a probabilidade de sucesso associada a um *output* binário dados os atributos de entrada. Formalmente, tem-se que:

$$p(x) = \Pr(y = 1|x). \quad (4.7)$$

sendo $x = (x^1, x^2, \dots, x^d)$ o vetor que representa uma coleção de variáveis independentes. Na [Equação 4.7](#), dado que y é numérica, o relacionamento entre $p(x)$ e x poderia ser modelado utilizando-se regressão linear simples, mas segundo [James et al. \(2013\)](#), sempre que uma linha reta é ajustada a uma resposta binária codificada como 0 e 1, em princípio pode-se obter $p(x) < 0$ para alguns valores de x e $p(x) > 1$ para outros, ou seja, algumas estimativas ficariam fora do intervalo $[0, 1]$, o que as torna difíceis de serem interpretadas como probabilidades. Para evitar este problema, é preciso modelar $p(x)$ utilizando-se uma função que fornece saídas compreendidas entre 0 e 1 para todos os valores de x . Antes de abordar detalhes do modelo logístico, é importante examinar um de seus conceitos chave, o *odds* de um evento.

Muitos autores destacam que é natural e intuitivo considerar probabilidades como forma de quantificar as chances de um evento ocorrer. Automaticamente são associados números compreendidos no intervalo $[0, 1]$, com os extremos representando, respectivamente, a certeza de não ocorrência e a certeza de ocorrência do evento. Existem, no entanto, outras maneiras de se representar as chances de um fenômeno, uma das quais é conhecida por *odds*, medida amplamente utilizada de relacionamento entre duas variáveis dicotômicas. Por definição, e conforme mostra a [Equação 4.8](#), o *odds* de um evento é a razão entre o número esperado de

vezes que um evento ocorrerá pelo número esperado de vezes que não ocorrerá:

$$O = \frac{p(x)}{1 - p(x)}. \quad (4.8)$$

A [Tabela 11](#) mostra a relação entre *odds* e probabilidade:

Tabela 11 – Relação entre probabilidade e *odds*.

Probabilidade	<i>odds</i>
0,0001	0,0100%
0,001	0,0999%
0,01	0,9901%
0,1	9,0909%
1	50,0000%
10	90,9091%
100	99,0099%
1000	99,9001%
10000	99,9900%

É possível observar que um *odds* menor do que 1 corresponde a probabilidades abaixo de 50%, enquanto que um *odds* maior do que 1 corresponde a probabilidades maiores do que 50%. A [Equação 4.8](#), assim como probabilidades, é limitada inferiormente por 0. Não existe, no entanto, limitante superior para o *odds*, podendo assumir qualquer valor entre $[0, \infty]$. Valores de *odds* próximos a zero indicam probabilidades muito baixas e valores próximos de ∞ indicam probabilidades muito altas.

Como foi colocado anteriormente, um dos problemas decorrentes de modelar uma variável resposta dicotômica (ou policotômica) com regressão linear é o fato de que probabilidades possuem limitantes inferiores e superiores, e funções lineares são ilimitadas. Uma maneira de contornar esta dificuldade é transformar a probabilidade em *odds*, o que remove o limitante superior. Tomando o logaritmo natural do *odds*, o limitante inferior também é removido. Igualando o resultado a uma função linear das d variáveis explicativas, obtém-se o modelo logístico [4.9](#):

$$\log \left[\frac{p(x)}{1 - p(x)} \right] = \beta_0 + \beta_1 x^1 + \dots + \beta_d x^d. \quad (4.9)$$

em que $\beta_0, \dots, \beta_d$ são constantes desconhecidas e $p(x)$ a probabilidade de que $y = 1$, definida em [4.7](#). A expressão ao lado esquerdo da equação é usualmente referida como *logit* ou *log-odds*. Portanto, tem-se que:

$$\text{logito}(p(x)) = \beta_0 + \beta_1 x^1 + \dots + \beta_d x^d. \quad (4.10)$$

É possível observar que o modelo de regressão logística de 4.9 tem um logito que é linear em x , e isto é praticamente o mesmo que modelar o logito da proporção de sucessos como uma função linear dos preditores. Resolvendo a Equação 4.9 para $p(x)$, obtém-se uma equação equivalente:

$$p(x) = \frac{e^{(\beta_0 + \beta_1 x^1 + \dots + \beta_d x^d)}}{1 + e^{(\beta_0 + \beta_1 x^1 + \dots + \beta_d x^d)}}. \quad (4.11)$$

Tanto a Equação 4.9 quanto a Equação 4.11 representam o modelo logístico. A expressão 4.11 pode ser simplificada ao se dividir o numerador e o denominador pelo próprio denominador, originando a Equação 4.12, a qual tem a desejada propriedade de produzir valores entre 0 e 1 independente dos valores atribuídos às variáveis:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x^1 + \dots + \beta_d x^d)}}. \quad (4.12)$$

Segundo Hosmer e Lemeshow (2000), no modelo de regressão logística o coeficiente de inclinação representa a alteração no logaritmo do *odds* (Equação 4.9) correspondente a mudança de uma unidade na variável independente. A interpretação dos coeficientes é mais facilmente discutida nos casos em que o modelo possui apenas uma variável preditora. Quando se tem duas ou mais variáveis predictoras, Hosmer e Lemeshow (2000) afirmam que o ajuste de uma série de modelos univariados raramente provê uma análise adequada dos dados, já que as variáveis independentes são frequentemente associadas umas com as outras e podem ter diferentes distribuições de acordo com as categorias da variável dependente. Portanto, considera-se uma análise multivariada para uma modelagem mais compreensiva dos dados e um dos objetivos de uma análise como esta é ajustar estatisticamente o efeito estimado de cada variável no modelo para diferenças nas distribuições e associações entre as outras variáveis independentes. Aplicando este conceito, que é chave para um claro entendimento das estimativas dos coeficientes de um modelo logito multivariado, pode-se supor que cada coeficiente provê uma estimativa do logaritmo do *odds*. Vale destacar, no entanto, que neste trabalho não é primário o interesse acerca da cuidadosa interpretação dos coeficientes estimados no modelo de regressão logística. Mais detalhes sobre esse assunto podem ser encontrados em (ref do livro de Hosmer!).

James *et al.* (2013) declaram que para estimar os parâmetros da Equação 4.9, um método chamado *maximum likelihood* é empregado. De maneira geral, este método provê estimativas para os parâmetros desconhecidos tais que a probabilidade predita $p(x)$ para um objeto individual corresponda o máximo possível ao valor observado da variável de interesse. Em outras palavras, procura-se encontrar parâmetros tais que, ao conectar suas estimativas no modelo para $p(x)$, obtém-se um número tão próximo quanto 1 para todos os casos de sucesso da base e um número tão próximo quanto 0 para os casos de fracasso. Para aplicar este método, é preciso construir

uma função chamada *likelihood function*, dada pela [Equação 4.13](#):

$$\ell(\beta) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})). \quad (4.13)$$

As estimativas β são escolhidas de forma a maximizar a expressão [4.13](#). Este método, segundo [James et al. \(2013\)](#), é uma abordagem utilizada para ajustar muitos modelos não-lineares. No contexto de regressão linear, o método de mínimos quadrados é de fato um caso especial do *maximum likelihood*, cujos detalhes estão fora do escopo deste trabalho e podem ser encontrados no livro de [Hosmer e Lemeshow \(2000\)](#).

A regressão logística é um método simples, que produz bons resultados e é amplamente utilizada em instituições financeiras. É um método afetado por variáveis redundantes, principalmente, mas que normalmente são removidas em uma etapa de seleção de variáveis. Além disso, a interpretação de atributos independentes nominais com muitas categorias pode ser difícil, dado que um coeficiente será estimado para cada uma delas e normalmente procura-se interpretá-los. Apesar de aumentar o custo de processamento, o método aceita muitos atributos independentes.

Métodos Forward, Backward e Stepwise

Dada a tarefa de selecionar um subconjunto de atributos preditores, o ideal seria realizar a seleção testando muitos modelos distintos, cada um deles contendo um grupo diferente de variáveis explicativas, e o melhor deles pode ser selecionado através de estatísticas que julgam a qualidade dos modelos, como AIC - *Akaike Information Criterion*, BIC - *Bayesian Information Criterion* e R^2 ajustado. Segundo James et. al (2013), a inviabilidade de testes assim reside no fato de que existem 2^d modelos contendo os possíveis subconjuntos de d variáveis. Se $d = 10$, então $2^{10} = 1024$ modelos deveriam ser considerados - em problemas práticos é muito comum deparar-se com mais de 10 variáveis explicativas. A menos que p seja muito pequeno, portanto, o processo de considerar todos os possíveis modelos torna-se impraticável e uma abordagem mais prática e eficiente faz-se necessária para determinar um conjunto menor de variáveis, e consequentemente de modelos. Ainda segundo James et. al (2013), existem três abordagens clássicas para esta tarefa:

- *Forward Selection*: inicia-se com o modelo nulo, que contém somente o intercepto. Após isto, d regressões logísticas são ajustadas e a variável que resulta no RSS mais baixo é adicionada ao modelo. Em seguida, a variável que resulta no RSS mais baixo para o novo modelo de duas variáveis é adicionada. Este procedimento é realizado até que algum critério de parada seja atingido.
- *Backward Selection*: inicia-se com um modelo contendo todas as variáveis e a variável com o maior valor- p , ou seja, a menos estatisticamente significativa, é removida. O novo modelo com $d - 1$ variáveis é ajustado e novamente a variável com o maior valor- p é

removida. Este procedimento é realizado até que algum critério de parada seja atingido ou até que os valores- p das variáveis restantes estejam abaixo de um valor determinado.

- *Stepwise Selection*: esta abordagem é uma mistura das abordagens anteriores. Inicia-se também com o modelo nulo e a variável que fornece o melhor ajuste é adicionada. À medida que as variáveis são adicionadas ao modelo, os valores- p das variáveis já selecionadas podem se tornar maiores. Portanto, se em determinado ponto o valor- p das variáveis no modelo exceder um nível determinado, então esta variável é removida do modelo. Este procedimento é executado até que todas as variáveis no modelo tenham um valor- p suficientemente baixo, e todas as variáveis fora do modelo tenham um valor- p grande se fossem adicionadas.

A abordagem *backward* não pode ser selecionada caso $d > n$, enquanto que a abordagem *forward* pode ser usada sempre. Esta, no entanto, pode incluir variáveis que, mais tarde, tornam-se redundantes. A abordagem *stepwise* pode evitar que isso aconteça.

4.4 Random Forests

Para tornar possível a discussão sobre *Random Forest*, é fundamental abordar uma classe de métodos baseados em regras de decisão, as quais são utilizadas para segmentar ou estratificar o espaço de variáveis preditoras em regiões distintas e não justapostas. Este procedimento pode ser representado por uma estrutura de árvore e é utilizado tanto em problemas de regressão quanto de classificação. Se a variável resposta é qualitativa, o método é chamado de árvore de classificação e, se a variável resposta é quantitativa, o método é chamado de árvore de regressão. Com o objetivo de fazer predição para um dado objeto do conjunto de dados, a média ou moda do treinamento da região do espaço d -dimensional à qual este objeto pertence é utilizada como estimativa. Assim sendo, objetos pertencentes a uma mesma região receberão estimativas iguais.

Formalmente, uma árvore de decisão é um grafo acíclico direcionado em que cada nó ou é um nó de divisão, com dois ou mais sucessores, ou um nó folha (FACELI *et al.*, 2011). Um nó folha, segundo os mesmos autores, é rotulado com uma função e usualmente compreende apenas os valores da variável resposta. Um nó de divisão, por sua vez, contém um teste condicional baseado nos valores das variáveis independentes. A Figura 16 exibe uma árvore de decisão e o correspondente espaço particionados por atributos x^1 e x^2 .

Como destacam James *et al.* (2013), computacionalmente não é viável considerar toda possível partição do espaço de variáveis preditoras e, por esta razão, uma abordagem conhecida como divisão recursiva binária, ou *recursive binary splitting*, é utilizada. Esta é uma abordagem em que a divisão inicia no topo da árvore e a melhor partição é escolhida a cada etapa do processo. Muitas medidas podem ser utilizadas para determinar a melhor partição, e elas são definidas em termos da distribuição das classes antes e após a divisão. Em um problema binário, a distribuição

Figura 16 – Árvore de decisão (à esquerda) e as regiões de decisão no espaço de atributos x^1 e x^2 .

Fonte: Faceli *et al.* (2011).

das classes em qualquer nó, ou partição, pode ser escrita como (p_0, p_1) , onde $p_1 = 1 - p_0$ e $p(c|t)$, ou simplesmente $p(c)$, denota a fração de observações pertencentes à classe c em uma dada partição t . A distribuição das classes determina o grau de impureza dos nós, de modo que quanto menor o grau de impureza, mais distorcida é a distribuição. Um nó com distribuição $(0, 1)$, por exemplo, têm impureza zero, enquanto que um nó com classes uniformemente distribuídas têm a maior impureza possível. Dentre as medidas de impureza, encontram-se:

$$\text{Entropia}(t) = - \sum_{c=1}^l p(c) \log_2 p(c). \quad (4.14)$$

$$\text{Gini}(t) = 1 - \sum_{c=1}^l [p(c)]^2. \quad (4.15)$$

$$\text{Erro de Classificação} = 1 - \max_c [p(c)]. \quad (4.16)$$

em que l é o número de classes e $0 \cdot \log_2 0 = 0$ no cálculo da entropia. O índice de Gini, segundo James *et al.* (2013), é uma medida de variância total dentre as classes e assume valores pequenos se as proporções $p(c)$ são próximas de 0 ou 1, indicando que um nó contém predominantemente observações de uma única classe. A entropia, por sua vez, toma valores próximos de 0 se as proporções $p(c)$ são próximas de 0 ou 1, indicando que, assim como o índice de Gini, esta medida assumirá valores pequenos se um nó é puro. De fato, e como mostra a Figura 17, as três medidas acima definidas são numericamente similares, pois atingem seus valores máximos quando a distribuição das classes é uniforme, ou seja, quando $p(c) = 0,5$, e atingem seus valores mínimos quando todas as observações pertencem à mesma classe, ou seja, quando $p(c) = 0$ ou $p(c) = 1$. James *et al.* (2013) declaram que o índice Gini ou a entropia são tipicamente utilizados para avaliar a qualidade de uma determinada divisão, já que estas duas medidas são

mais sensíveis ao grau de pureza do que é a taxa de erro de classificação.

Figura 17 – Comparação das medidas de impureza para problemas de classificação binária.

Fonte: Faceli *et al.* (2011).

De acordo com a definição de Faceli *et al.* (2011), no contexto de árvore de decisão a entropia é usada para medir a aleatoriedade o atributo alvo. O atributo que mais reduz a aleatoriedade da variável resposta, como afirmam os autores, será escolhido para dividir os dados a cada nó de decisão. Em outras palavras, para determinar o desempenho de uma variável em uma determinada quebra, a entropia do nó que antecede a divisão é comparada à entropia do nó que sucede a divisão e, quanto maior a diferença, melhor a condição de teste. Uma medida conhecida como *Information Gain*, representada pela Equação 4.17, é um critério utilizado para mensurar a redução na entropia nas partições obtidas de acordo com os valores da variável:

$$IG(A) = I(t) - \sum_{i=1}^v \frac{N(t_v)}{N(t)} \cdot I(t_v). \quad (4.17)$$

sendo $IG(A)$ o ganho de informação associado ao atributo selecionado A , $I(t)$ a medida de impureza associada ao nó t que antecede a divisão, $N(t)$ o total de observações no nó t , v o índice que percorre os diferentes valores de A escolhidos para dividir os dados e $N(t_v)$ e $I(t_v)$ o número de observações e a medida de impureza associados aos nós que sucedem a divisão, respectivamente. Um algoritmo de árvore de decisão conhecido como C4.5 faz uso de condições de teste que maximizam $IG(A)$ a cada quebra. Dado que a impureza do nó que antecede a

divisão é a mesma para qualquer que seja o atributo selecionado, maximizar 4.17 é o mesmo que minimizar a média ponderada da medida de impureza dos nós que sucedem a divisão. Um outro algoritmo, conhecido como CART, utiliza o índice de Gini como critério de seleção de atributos. Neste caso, como afirmam Faceli *et al.* (2011), a impureza média ponderada dos nós que sucedem a divisão é subtraída do valor da impureza do nó que antecede a divisão, e o atributo que resulta na maior diminuição é selecionado.

A divisão recursiva pode continuar indefinidamente e gerar árvores complexas que produzem boas previsões no conjunto de treinamento, mas resultados não tão bons no conjunto de teste. Para James *et al.* (2013), uma possível alternativa é construir a árvore até o ponto em que o decréscimo de uma medida de impureza de uma determinada partição não exceda um limiar estabelecido, o que resultará em árvores menores, com menos divisões, mas é possível que o ponto em que houve a parada, ou poda, anteceda uma boa quebra. Segundo Faceli *et al.* (2011), esta abordagem é conhecida como pré-poda e outras regras de parada podem ser adotadas, como estipular um número mínimo de observações em um nó ou determinar que todas as observações em um nó pertençam a uma mesma classe.

Uma outra estratégia de parada, comumente utilizada e conhecida como pós-poda, é construir uma árvore completa T_0 , superajustada aos dados do treinamento, e depois podá-la, gerando uma subárvore T . A decisão chave na pós-poda é podar ou não uma subárvore (FACELI *et al.*, 2011). Uma maneira de fazer essa escolha, como colocam James *et al.* (2013), é considerar uma sequência de árvores indexadas por um parâmetro não negativo α , o qual pondera o tamanho da subárvore em relação à taxa de erro de treinamento. Para cada valor deste parâmetro há uma subárvore correspondente $T \subset T_0$ tal que

$$R_\alpha(T) = R(T) + \alpha|T|. \quad (4.18)$$

é tão pequeno quanto possível. Na Equação 4.18, $|T|$ indica o número de nós terminais da árvore T e $R(T)$ é a taxa de erro. Se $\alpha = 0$, a subárvore T será simplesmente igual a T_0 e, à medida que α cresce, a quantidade 4.18 tende a ser minimizada e a subárvore terá poucos nós terminais. Faceli *et al.* (2011) resumem este como sendo um processo que gera um número finito de subárvores com nós terminais progressivamente menores e a árvore podada é aquela que minimiza $R_\alpha(T)$ na sequência de subárvores. Esta metodologia é conhecida como *custo complexidade* e o valor do parâmetro α pode ser determinado em um conjunto de validação ou através de validação cruzada.

Segundo James *et al.* (2013), árvores de decisão sofrem de alta variância. Uma maneira natural de reduzi-la é gerar amostralmente B subconjuntos de treinamento a partir do conjunto de dados original, procedimento conhecido como *bootstrap*, treinar então um modelo em cada

subconjunto b , a fim de obter $\hat{f}^b(x)$ e, finalmente, calcular a média das predições:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(x). \quad (4.19)$$

Esta abordagem é conhecida como *bagging* e é especialmente útil para árvores de decisão. Para aplicá-la à árvores de classificação, pode-se registrar a classe predita por cada uma das árvores e obter uma votação majoritária, no sentido de que a predição geral é a classe mais comumente encontrada entre as previsões dos B subconjuntos. O número de árvores B , ainda segundo os autores, não é um parâmetro crítico. Na prática, adota-se um número suficientemente grande para que o erro se estabeleça - um número grande de B não ocasionará *overfitting*.

James *et al.* (2013) afirmam que há uma maneira direta de estimar o erro de teste de um modelo *bagging*, sem a necessidade de realizar validação cruzada ou utilizar um conjunto apartado de validação. É possível mostrar que ao gerar amostras de treinamento a partir do conjunto de dados original, somente cerca de dois terços das observações são consideradas, e isto ocorre porque os subconjuntos têm tamanho igual ao da população e são gerados a partir de um processo de amostragem com reposição. As observações não utilizadas para ajustar uma determinada árvore, e que fazem parte do terço não considerado, são referidas como observações *out-of-bag* (OOB) e pode-se prever a variável resposta associada à estas observações utilizando-se as árvores das quais não fizeram parte. Com o objetivo de obter uma única predição, o voto majoritário é tomado e, após isto, o erro de classificação, conhecido como erro OOB, é calculado. O erro OOB resultante é uma estimativa válida do erro de teste para o modelo *bagging*, uma vez que a resposta para cada observação é prevista usando apenas as árvores que não foram ajustadas usando essa observação (JAMES *et al.*, 2013).

Random Forests, ou florestas aleatórias, apresentam uma melhoria em relação à abordagem *bagging* por meio da inserção de um ajuste que procura reduzir a correlação entre as árvores. Neste método, muitas árvores de decisão também são construídas a partir de subconjuntos dos dados de treinamento, mas diferentemente do método *bagging*, a cada divisão uma amostra aleatória de variáveis preditoras apenas é considerada. Ou seja, para cada particionamento realizado, considera-se como candidatos $a < d$ atributos independentes e, tipicamente, $a \approx \sqrt{d}$. Este racional pode ser explicado pelo fato de que, caso haja uma variável com forte poder preditivo no conjunto de dados e uma série de outras variáveis cujo poder preditivo é moderado, na coleção de árvores ajustadas uma será semelhante à outra - o atributo forte provavelmente aparecerá em todas as árvores. Assim, as predições advindas destas árvores serão também correlacionadas e, como declaram James *et al.* (2013), calcular a média de valores correlacionados não leva a tão grande redução na variância. As florestas aleatórias superam este problema, forçando cada divisão a considerar apenas um subconjunto de preditores. A principal diferença, portanto, entre *bagging* e *random forests* é a escolha do tamanho do subconjunto de variáveis preditoras. Se $a = d$, então as duas abordagens produzirão resultados iguais.

Apesar de ser um método derivado de árvores de decisão, o grau de interpretabilidade de uma floresta aleatória é reduzido devido ao número de árvores que a compõem. Neste sentido, não é possível exibir um único modelo produzido ao final do processo, já que o resultado final é gerado a partir da contribuição individual de cada árvore. Ao passo que a interpretabilidade é prejudicada, o grau de acurácia aumenta, a natureza aleatória da construção de cada árvore minimiza o sobreajuste, o método consegue lidar com conjuntos de dados com muitos objetos e/ou atributos, possibilita a substituição de dados ausentes ou cria uma estimativa para este grupo e também gera métricas de importância de variáveis.

4.5 Redes Neurais

O desenvolvimento de redes neurais artificiais (RNAs) foi inspirado na estrutura e funcionamento do sistema nervoso, formado por um conjunto extremamente complexo de células nervosas, os neurônios, responsáveis por conduzir os impulsos nervosos através do corpo. Os neurônios são conhecidos como a unidade básica do sistema nervoso e a comunicação entre eles ocorre através das sinapses, regiões ativas de contato que permitem a transmissão de sinais nervosos. Segundo [Mitchell \(1997\)](#), estima-se que o cérebro humano possua uma rede densamente interconectada de aproximadamente 10^{11} neurônios, cada um conectado, em média, a 10^4 outros. Esta estrutura biológica emaranhada confere aos seres humanos capacidade singular no processamento de diversas informações e execução de múltiplas tarefas, das mais simples às mais complexas. O objetivo das RNAs é simular a capacidade de aprendizado do cérebro humano através de técnicas computacionais e modelos matemáticos.

O Modelo Matemático de McCulloch e Pitts

Em 1943 o Psiquiatra e Neuroanatomista McCulloch e o Matemático Pitts propuseram um modelo matemático de um sistema neural em que as unidades básicas funcionam de maneira simples, mas que produzem um elevado poder computacional quando conectadas. As componentes do modelo de McCulloch e Pitts são descritas a seguir:

- Conjunto de pesos $w_{j\kappa}$, correspondentes às sinapses, que conectam um sinal ou unidade de entrada x^j ao neurônio κ , cuja atividade é binária - o neurônio dispara ou não dispara;
- Um somatório $\sum$ que totaliza os sinais de entradas já ponderados pelos pesos $w_{j\kappa}$. Pode ser equacionado como $v_{\kappa} = \sum_{j=1}^d w_{j\kappa} x^j$;
- Uma função de ativação $\phi(v_{\kappa})$ que define a saída y do neurônio tomando como base o valor gerado pelo somatório $\sum$. O neurônio de McCulloch e Pitts, como destaca [Marsland \(2014\)](#), emprega uma função de ativação limiar, representada pela [Equação 4.20](#). Quando a soma das entradas ponderadas pelos pesos $w_{j\kappa}$ está acima de um limiar estabelecido, o

resultado produzido será igual a 1, indicando que o neurônio dispara; caso contrário, será igual a 0, indicando que o neurônio não dispara.

$$o = \phi(v_k) = \begin{cases} 1, & v_k > \theta \\ 0, & v_k \leq \theta \end{cases} \quad (4.20)$$

A função limiar é utilizada no modelo matemático de McCulloch e Pitts, mas muitas funções têm sido utilizadas em RNAs e propostas na literatura, a exemplo das funções linear e sigmoidal, a qual será apresentada mais adiante.

A Figura 18 retrata o modelo proposto por McCulloch e Pitts. À direita da figura estão representados os elementos de entrada x^j , $j = 1, \dots, d$, em que d representa a dimensão da base de dados. Cada objeto x_i da base de dados é apresentado ao neurônio, ou seja, as entradas $x^1, \dots, x^j$ são correspondentes a um objeto i qualquer da base de dados, $i = 1, \dots, n$. Em neurônios reais, os elementos de entrada representam os sinais advindos de outros neurônios. Um sinal de entrada x^j conectado ao neurônio k é multiplicado pelo peso sináptico w_{jk} . Estas conexões podem ser excitatórias ($w_{jk} > 0$) ou inibitórias ($w_{jk} < 0$), o que torna os neurônios mais ou menos propensos a disparar, respectivamente. No modelo de McCulloch e Pitts, a chegada de uma única sinapse inibitória evita absolutamente o disparo do neurônio, independente do número de sinapses excitatórias que estejam chegando conjuntamente com a sinapse inibitória.

Figura 18 – Modelo matemático de um neurônio proposto por McCulloch e Pitts (1943).

Fonte: <<http://conteudo.icmc.usp.br/pessoas/andre/research/neural/>>.

Os pesos sinápticos do modelo de McCulloch e Pitts assumem valores constantes, e isto significa que se alimentarmos o neurônio com os mesmos dados de entrada repetidamente, a saída será sempre a mesma, ou seja, ele sequer aprende - sua atividade se restringe tão somente a disparar ou não. Para torná-lo mais atrativo, é preciso descobrir como fazê-lo aprender, e esta é a característica que diferencia as RNAs de um modelo simplista como o de McCulloch e Pitts. Partindo do pressuposto de que uma tarefa é desconhecida em um ponto de partida t , para que o aprendizado ocorra é preciso que alguma mudança ou adaptação seja observada. Dos elementos

que compõem um neurônio, os pesos e o valor do limiar, ou *threshold*, podem ser adaptados de forma que o seu desempenho seja aperfeiçoado, e este é justamente o exercício a que se dedicam os algoritmos de treinamento de redes neurais. À propósito, [Marsland \(2014\)](#) destaca este como sendo um aspecto interessante, pois indica que o aprendizado ocorre essencialmente nos pesos, ou seja, entre os neurônios, e não exatamente nos neurônios.

O modelo de McCulloch e Pitts é uma simplificação do que se sabia na época sobre o neurônio biológico, mas não teve muita influência na neurobiologia, pois as evidências experimentais que começavam a se acumular a partir da década de 1940 indicavam que o neurônio é mais complexo do que um simples elemento binário com um limiar. Apesar de simples, o modelo gerou impacto na ciência da computação e serviu de base para o desenvolvimento de redes neurais mais complexas. Segundo [Faceli et al. \(2011\)](#), outros pesquisadores também contribuíram com o tema, como Hebb (1949), com estudos sobre aprendizado, e Rosenblatt (1958), com a teoria sobre os perceptrons. A publicação do livro de Minsky e Papert (1969), no qual os autores apontaram a limitação da rede perceptron a problemas linearmente separáveis, foi um dos motivadores para a interrupção das pesquisas sobre RNAs. Na década de 1980, no entanto, o surgimento de computadores mais rápidos fez ressurgir interesse pelo assunto.

A Rede Perceptron

A primeira rede neural a ser implementada, segundo [Faceli et al. \(2011\)](#), foi a rede perceptron, desenvolvida por Rosenblatt em 1958. De acordo com a definição de [Gunther e Fritsch \(2010\)](#), o perceptron ([Figura 19](#)) consiste de uma camada de entrada com d covariáveis, também conhecidas como variáveis de entrada, e uma camada de saída com um neurônio que produz os *outputs* da rede. A camada de entrada inclui um elemento constante x^0 , normalmente igual a -1 , referente às sinapses que não são diretamente influenciadas por nenhuma covariável. O bias, como é designado, representa uma maneira de contornar o problema ocasionado por um vetor de entrada com elementos exclusivamente nulos, o que produziria um resultado sempre igual a 0 (zero) independente do valor dos pesos sinápticos. Uma estrutura de rede tal como descrita calcula o *output* segundo a função representada pela [Equação 4.21](#):

$$y_{\kappa}(x) = \phi \left(\sum_j w_{j\kappa} x^j \right). \quad (4.21)$$

em que $x = (x^0, \dots, x^d)$ representa o vetor de todas as covariáveis, incluindo o bias, e $w = (w_0, \dots, w_d)$ o vetor de todos os pesos sinápticos, os quais determinam a contribuição da entrada x à saída da rede. Conforme a notação já apropriada na descrição dos elementos do modelo de McCulloch e Pitts, o índice κ faz referência ao neurônio ao qual os elementos do vetor x estão conectados, e pode ser especialmente útil nos casos em que a rede perceptron possui mais de um neurônio e eventualmente mais de uma camada.

Figura 19 – Rede perceptron.

Fonte: <<http://novos-cientistas.blogspot.com/2012/05/redes-neurais.html>>.

Para Haykin (2001), o perceptron é a forma mais simples de uma rede neural normalmente usada para classificação de padrões ditos *linearmente separáveis* e consiste basicamente de um único neurônio com pesos sinápticos ajustáveis e bias. O perceptron construído em torno de um único neurônio, como afirma o autor, é limitado a realizar classificação de padrões com apenas duas classes, mas expandindo a camada de saída do perceptron para incluir mais de um neurônio é possível realizar classificação com mais de duas classes.

A avaliação do erro cometido pela rede perceptron para um objeto x_i ocorre por meio da comparação das *targets* que se têm disponíveis com os *outputs* binários produzidos, como mostra a Equação 4.22:

$$E = \sum_{i=1}^n (t_i - y_i). \quad (4.22)$$

em que y_i é a saída produzida pela rede e t_i é o valor observado da variável resposta. Naturalmente, o que se quer evitar é que um neurônio dispare quando não deveria disparar ou que não dispare quando deveria disparar, ou seja, que produza um *output* 1 quando a *target* é 0 ou vice-versa. Caso uma destas situações ocorra, então os pesos devem ser ajustados para que o neurônio acerte da próxima vez. O ajuste dos pesos relativos a uma unidade de entrada fixa x^l e uma unidade de saída também fixa κ ocorre durante o treinamento segundo a Equação 4.23:

$$w_{\kappa l} = w_{\kappa l} + \eta (t_l - y_l) x^l. \quad (4.23)$$

sendo $w_{\kappa l}$ o peso que conecta a unidade de entrada l ao neurônio κ , η uma taxa que representa a

velocidade de aprendizado da rede, t_i a saída observada e y_i a saída produzida pela rede. A taxa de aprendizagem η controla a quantia pela qual os ajustes nos pesos são realizados. Valores altos fazem com que os pesos mudem muito sempre que uma resposta errada é produzida, o que torna a rede instável; taxas menores, por sua vez, tornam a rede mais estável e resistente aos ruídos, mas aumentam o custo computacional já que os pesos precisarão ver com mais frequência as entradas da rede antes de mudar significativamente. Marsland (2014) sugere que sejam utilizados valores moderados ($0,1 < \eta \leq 0,4$). Além disso, vale ressaltar que os pesos serão ajustados conforme a equação acima para cada elemento j do vetor de entrada.

Segundo Mitchell (1997), a rede perceptron pode ser vista como uma superfície de decisão de hiperplanos no espaço n -dimensional de objetos (ou pontos). Rosenblatt provou que se os objetos usados para treinar o perceptron são tirados de duas classes linearmente separáveis, então o algoritmo do perceptron converge e posiciona a superfície de decisão na forma de um hiperplano entre as duas classes (HAYKIN, 2001). Certamente, nem todos os conjuntos de números positivos e negativos podem ser separados por um hiperplano. Portanto, embora a rede perceptron seja eficiente na resolução de problemas lineares, ela pode não convergir se os exemplos não forem linearmente separáveis (Figura 20). À exemplo disto, Mitchell (1997) destaca que perceptrons podem ser usados para representar todas as funções booleanas primitivas AND, OR, NAND e NOR, mas funções como XOR, por exemplo, não podem ser representadas por um único perceptron. Visto que redes de unidades limiares podem representar uma vasta variedade de funções e unidades isoladas não podem, geralmente há maior interesse em redes multicamadas.

Figura 20 – Conjunto de exemplos de treinamento separáveis linearmente (à esquerda) e não separáveis linearmente (à direita).

Fonte: <<http://www.eletrica.ufpr.br/ufpr2/professor/36/TE808/7-RNA-AM.pdf>>.

Redes Multi-layer Perceptron e o Algoritmo Backpropagation

Para resolver problemas não linearmente separáveis utilizando RNAs, a alternativa mais utilizada é adicionar uma ou mais camadas intermediárias (FACELI *et al.*, 2011). Cybenko (1989 apud FACELI *et al.*, 2011, p. 115) afirma que uma rede com uma camada intermediária pode implementar qualquer função contínua. Marsland (2014) declara que duas camadas ocultas é o máximo que se precisa para resolver problemas, e ainda afirma ser possível mostrar matematicamente que uma camada com muitos nós ocultos (neurônios) é o suficiente para o aprendizado de uma rede multicamadas. A má notícia, segundo o autor, é que não há teoria para orientar a escolha do número de nós ocultos, devendo isto ser feito por meio de experimentação.

Segundo a definição de Haykin (2001), as redes de múltiplas camadas (Figura 21) são uma importante classe de redes neurais e consistem: (1) de um conjunto de nós de fonte que compõem a camada de entrada, (2) uma ou mais camadas ocultas de nós computacionais e (3) uma camada de saída de nós computacionais. Estas redes representam uma generalização do perceptron de camada única e por isso são normalmente chamadas de perceptron de múltiplas camadas ou *multi-layer perceptron* (MLP).

Figura 21 – Rede neural consistindo de múltiplas camadas de neurônios conectados (MLP).

Fonte: <[https://www.researchgate.net/figure/](https://www.researchgate.net/figure/Rede-multicamadas-Uma-rede-multilayer-perceptron-e-uma-rede-progressiva-ou-seja-uma-RNA_fig2_282354973)

Rede-multicamadas-Uma-rede-multilayer-perceptron-e-uma-rede-progressiva-ou-seja-uma-RNA_fig2_282354973>.

Marsland (2014) descreve o treinamento de redes multicamadas em duas etapas, comumente chamadas de *forward* e *backward*, e que consistem em calcular as saídas de acordo com

as entradas e os pesos atuais e, em seguida, atualizar os pesos de acordo com o erro. Na fase *forward*, cada objeto x_i apresentado à rede é recebido por cada um dos neurônios na camada intermediária e seus valores são então ponderados pelos respectivos pesos e posteriormente somados. Após isto, a função de ativação é aplicada, um valor de saída é produzido e utilizado como entrada pelos neurônios da camada seguinte. Esse processo continua até que um valor seja produzido pelo neurônio na camada de saída, quando é então comparado ao valor observado da variável de interesse referente àquele objeto. Conforme equacionado em [Gunther e Fritsch \(2010\)](#), um procedimento tal como descrito, considerando uma única camada oculta, calcula o *output* segundo as equações abaixo:

- O cálculo da ativação de um neurônio ζ na camada oculta é dado pela [Equação 4.24](#):

$$\begin{aligned} a_{\zeta}(x) &= g(v_{\zeta}) \\ &= g\left(\sum_j x^j \cdot v_{j\zeta}\right). \end{aligned} \quad (4.24)$$

em que $a_{\zeta}(x)$ representa o resultado da ativação do neurônio oculto ζ para o vetor de entrada x , j é o índice que percorre os elementos do vetor de entrada, g é a função de ativação da rede e $v_{j\zeta}$ é o peso entre a unidade oculta ζ e o j -ésimo elemento de entrada.

- O cálculo da ativação de um neurônio κ na camada de saída é dado pela [Equação 4.25](#):

$$\begin{aligned} y_{\kappa}(x) &= g(v_{\kappa}) \\ &= g\left(\sum_l a_l \cdot w_{l\kappa}\right) \\ &= g\left(\sum_l w_{l\kappa} \cdot g\left(\sum_j v_{jl} \cdot x^j\right)\right). \end{aligned} \quad (4.25)$$

em que l é o índice que percorre os neurônios da camada oculta, a_l é a ativação do neurônio oculto l e $w_{l\kappa}$ é o peso entre a ativação oculta a_l e a unidade de saída κ .

Prosseguindo com o treinamento da rede, o valor do erro na camada de saída é calculado e, em seguida, utilizado na fase *backward* para ajustar seus pesos, e o ajuste prossegue da camada de saída até a camada de entrada. É importante ressaltar, no entanto, que as ativações de entrada não são conhecidas para os neurônios na camada de saída e as saídas não são conhecidas para neurônios na camada oculta - caso a rede possua mais de uma camada oculta, nem as entradas e nem as saídas são conhecidas. Por isso, como afirmam [Faceli et al. \(2011\)](#), os erros são conhecidos apenas para os neurônios da camada de saída, e precisam ser estimados para os neurônios das camadas intermediárias.

Um obstáculo que havia para utilizar redes multicamadas era a ausência de um algoritmo para o treinamento destas redes, o que foi transposto com a proposta de um algoritmo baseado em gradiente descendente denominado *backpropagation* (FACELI *et al.*, 2011). Este algoritmo emprega gradiente descendente para minimizar o erro quadrado entre as saídas produzidas pela rede e os valores observados, as *targets*. Para tanto, as derivadas parciais da função do erro, dada pela Equação 4.26, são calculadas em relação aos conjuntos de pesos de cada camada, e a direção negativa do gradiente é norteadora nesse processo. O desenvolvimento das derivadas, calculadas utilizando-se a regra da cadeia, dá origem tanto à equação de atualização dos pesos quanto às equações de cálculo do erro na camada de saída e na camada oculta. Marsland (2014) afirma que a regra da cadeia é útil, pois permite escrever as ativações dos neurônios de saída em termos das ativações dos neurônios ocultos e dos pesos da camada de saída, e então o erro pode ser enviado de volta através da rede para a camada oculta para decidir quais são as saídas para estes neurônios.

$$E(v, w) = \frac{1}{2} \sum_{i=1}^n (t_i - y_i)^2. \quad (4.26)$$

Os erros e as atualizações dos pesos de uma rede Multi-layer perceptron são calculados segundo as equações abaixo:

- O erro da camada de saída o associado ao neurônio κ é calculado da seguinte maneira:

$$\delta_o(\kappa) = y_\kappa (1 - y_\kappa) (t_\kappa - y_\kappa), \quad (4.27)$$

em que t_κ é a saída observada e y_κ a saída produzida pela rede para um neurônio fixo κ .

- Para uma unidade de saída κ e unidade oculta ζ , a atualização dos pesos na camada de saída é dada por:

$$w_{\zeta\kappa} = w_{\zeta\kappa} + \eta \delta_o(\kappa) a_\zeta, \quad (4.28)$$

em que a_ζ representa a saída de um neurônio oculto ζ e $w_{\zeta\kappa}$ representa o peso entre a saída do neurônio oculto ζ e o neurônio de saída κ .

- O erro da camada oculta h associado a um neurônio fixo ζ é calculado da seguinte maneira:

$$\delta_h(\zeta) = a_\zeta (1 - a_\zeta) \sum_k w_{\zeta k} \delta_o(k), \quad (4.29)$$

em que a_ζ é a saída do neurônio oculto ζ e $w_{\zeta k}$ o peso que conecta a saída do neurônio oculto ζ ao k -ésimo neurônio da camada de saída.

- Para uma unidade oculta fixa ζ e unidade de entrada também fixa ι , a atualização de pesos na camada oculta é dada por:

$$v_{\iota\zeta} = v_{\iota\zeta} - \eta \delta_h(\zeta) x^\iota, \quad (4.30)$$

em que $v_{i\zeta}$ é o peso que conecta a entrada i ao neurônio oculto ζ e x^i é o atributo de entrada i referente ao i -ésimo objeto.

As fórmulas acima valem para redes com uma camada extra entre as camadas de entrada e saída, mas vale ressaltar que pode-se estender estes cálculos à redes com mais de uma camada oculta.

Redes multicamadas utilizam nas camadas intermediárias funções de ativação não lineares, a exemplo da função sigmoideal, cuja forma é similar à da função limiar, mas guarda o importante aspecto de ser diferenciável. O algoritmo *backpropagation* requer, por motivos claros, que a função de ativação seja contínua e diferenciável, exigências estas atendidas pela função sigmoideal. No algoritmo acima descrito, neurônios sigmoideais foram utilizados, e a derivada da função de ativação em questão também compõe a [Equação 4.28](#) e a [Equação 4.30](#) de cálculo dos erros. Segundo [Marsland \(2014\)](#), a forma mais comumente utilizada dessa função, onde o parâmetro β é positivo, é:

$$a_\zeta = g(v_\zeta) = \frac{1}{1 + \exp(-\beta v_\zeta)}. \quad (4.31)$$

A função representada pela [Equação 4.31](#) é útil em problemas de classificação, pois retorna valores no intervalo $[0, 1]$, mas para problemas de regressão ela pode não ser tão interessante. Neste caso, outras funções podem ser empregadas.

[Mitchell \(1997\)](#) destaca que uma das maiores diferenças no caso de redes multicamadas é que a superfície de erro pode ter múltiplos mínimos locais, e isto significa que a convergência do método gradiente descendente é garantida para um mínimo local, e não necessariamente um mínimo global. Diferentemente das redes perceptron, não existe um teorema de convergência para algoritmos de aprendizado para o treinamento de redes multicamadas ([FACELI *et al.*, 2011](#)). Uma das críticas feita ao algoritmo é a lentidão na convergência para um bom conjunto de pesos e a queda de desempenho quando dados massivos são utilizados - o número de parâmetros ajustáveis aumenta se mais atributos forem adicionados à base e mais neurônios e camadas forem adicionados à arquitetura da rede. Apesar deste obstáculo, [Mitchell \(1997\)](#) afirma que o algoritmo produz resultados excelentes em muitas aplicações. Na tentativa de contornar este problema, [Marsland \(2014\)](#) sugere que sejam realizadas repetições de treinamento das redes de tal forma que diferentes pontos de partida, eventualmente mais próximos de um mínimo global, sejam testados.

Os ajustes dos pesos no algoritmo continuam até que seja atingido um critério de parada. [Faceli *et al.* \(2011\)](#) afirmam que diferentes critérios podem ser utilizados, como por exemplo, um número máximo de ciclos ou repetições ou uma taxa máxima de erro. O treinamento também pode ser encerrado quando a taxa de erro para o conjunto de validação começa a subir, e este processo é conhecido como validação cruzada com *early stop*.

Existem variações do algoritmo *backpropagation* bastante utilizadas e que, segundo [Gunther e Fritsch \(2010\)](#), também procuram minimizar a função do erro adicionando uma taxa de aprendizagem aos pesos. Para a variação do algoritmo conhecida como *resilient backpropagation*, uma taxa de aprendizagem η_k também é usada para cada conjunto de pesos e pode ser alterada durante o treinamento. Isto soluciona, segundo os autores, o problema de definir uma taxa de aprendizagem global adequada a todo o processo de formação da rede. Além disso, ao invés da magnitude das derivadas parciais, apenas seu sinal é usado para atualizar os pesos. [Gunther e Fritsch \(2010\)](#) ainda destacam que para acelerar a convergência em áreas rasas, a taxa de aprendizagem η_k será aumentada se a derivada parcial correspondente mantiver seu sinal e, pelo contrário, será reduzida se a derivada parcial da função do erro mudar seu sinal - a mudança de sinal indica que o mínimo da função é perdido devido a uma taxa de aprendizado muito grande.

O desenvolvimento da primeira etapa de treinamento, *foward*, pode ser representado matricialmente. Neste exemplo consideraremos uma rede com 1 camada oculta e 1 neurônio na camada de saída, suficiente para um problema de classificação binário, mas as etapas podem ser estendidas para redes mais complexas. Sejam a matriz $X = [x_{ij}]_{n \times d}$ dos exemplos de entrada e as matrizes $V = [v_{ij}]_{d \times u}$ e $W = [w_{ij}]_{u \times 1}$ de pesos da primeira camada e da camada de saída, respectivamente. O índice n se refere ao número de vetores de entrada, d ao número de elementos em cada vetor e u ao número de neurônios da primeira camada. O produto das matrizes X e V resulta em uma terceira matriz A de dimensões $n \times u$ que carrega o resultado do somatório de cada neurônio da camada oculta para cada vetor de x_i de entrada:

$$\begin{aligned}
 X \times V &= \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix}_{n \times d} \times \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1u} \\ v_{21} & v_{22} & \dots & v_{2u} \\ \vdots & \vdots & \ddots & \vdots \\ v_{d1} & v_{d2} & \dots & v_{du} \end{bmatrix}_{d \times u} = \\
 &= \begin{bmatrix} x_{11} * v_{11} + x_{12} * v_{21} + \dots + x_{1d} * v_{d1} & \dots & x_{11} * v_{1u} + x_{12} * v_{2u} + \dots + x_{1d} * v_{du} \\ x_{21} * v_{11} + x_{22} * v_{21} + \dots + x_{2d} * v_{d1} & \dots & x_{21} * v_{1u} + x_{22} * v_{2u} + \dots + x_{2d} * v_{du} \\ \vdots & \vdots & \vdots \\ x_{n1} * v_{11} + x_{n2} * v_{21} + \dots + x_{nd} * v_{d1} & \dots & x_{n1} * v_{1u} + x_{n2} * v_{2u} + \dots + x_{nd} * v_{du} \end{bmatrix}_{n \times u} = \\
 &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1u} \\ a_{21} & a_{22} & \dots & a_{2u} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nu} \end{bmatrix}_{n \times u} = A
 \end{aligned}$$

A matriz A será ainda multiplicada pela matriz dos pesos que conecta cada um dos neurônios ao *output*:

$$A \times W = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1u} \\ a_{21} & a_{22} & \dots & a_{2u} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nu} \end{bmatrix}_{n \times u} \times \begin{bmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{u1} \end{bmatrix}_{u \times 1} = \begin{bmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{bmatrix}_{n \times 1} = T$$

[Faceli et al. \(2011\)](#) afirmam que o constante aumento no número de aplicações que incorporam RNAs confirma a sua importância para a resolução de problemas. Uma das características que torna este método popular é a capacidade de generalização e tolerância a falhas e ruídos. A interpretabilidade e entedimento, no entanto, são duas grandes dificuldades. Entender como os resultados foram gerados é ter que desvendar uma caixa-preta, como é conhecido o método. A grande dificuldade, segundo [Faceli et al. \(2011\)](#), é que o conhecimento está armazenado na forma de uma grande quantidade de parâmetros, os quais são manipulados através de complicadas fórmulas matemáticas. A escolha pelo método é normalmente pautada nos bons resultados que eles produz.

AVALIAÇÃO DOS MODELOS PREDITIVOS

Este capítulo aborda um conjunto de técnicas cujo objetivo é avaliar o poder preditivo dos modelos de classificação. As medidas apresentadas podem ser utilizadas em problemas cuja variável resposta possui duas classes, ou seja, em problemas de classificação binários. Estas medidas serão utilizadas no [Capítulo 6](#) para avaliar e comparar os algoritmos estudados.

5.1 Amostragem

Conforme visto na [Subseção 1.1.1](#), para indução e avaliação dos classificadores o conjunto de dados total deve ser particionado em três subconjuntos disjuntos: treinamento, validação e teste. Métodos de amostragem podem ser utilizados para este fim, e segundo [Faceli et al. \(2011\)](#), os principais existentes são:

- *Holdout*: o conjunto de dados é dividido em uma porcentagem p para treinamento e $(1 - p)$ para teste. A porcentagem p para treinamento ainda pode ser dividida para obtenção do conjunto de validação. Normalmente as proporções 60% e 40% ou 70% e 30% são utilizadas. Uma das desvantagens deste método é que ele não permite avaliar o quanto o desempenho de uma técnica varia quando diferentes combinações de objetos são apresentadas em seu treinamento ([FACELI et al., 2011](#)). Ou seja, este método possui apenas uma iteração, dependendo muito da qualidade da partição escolhida. Uma maneira de contornar este problema é gerar partições aleatórias para compor os conjuntos, e depois tomar a média do desempenho, método este conhecido como amostragem aleatória.
- *Validação Cruzada ou K-Fold Cross Validation*: neste método, o conjunto de dados é particionado aleatoriamente em k subconjuntos e todos os subconjuntos participam do conjunto de treinamento e do conjunto de teste. O parâmetro k , portanto, define também o número de vezes em que medidas de desempenho serão geradas. A cada processamento, um dos subconjuntos é utilizado como conjunto de teste e os outros $k - 1$ subconjuntos são

utilizados para treinamento. Segundo [Mitchell \(1997\)](#), o valor $k = 10$ é tipicamente usado na literatura. Ao final, são tomados a média e o desvio padrão das medidas de desempenho. Uma das desvantagens desse método é o aumento do tempo computacional e a dificuldade em gerar k partições em conjuntos de dados pequenos. Por outro lado, segundo [Viana \(2015\)](#), é útil para verificar se uma pequena mudança na amostra provoca uma grande diferença nos parâmetros.

- *Bootstrap*: k subconjuntos de treinamento também são gerados, mas a partir de um processo aleatório com reposição. Potanto, um exemplo pode estar em um subconjunto mais de uma vez. Os exemplos não selecionados para os subconjuntos de treinamento farão parte do subconjunto de teste. Segundo [Faceli et al. \(2011\)](#), a intenção é repetir o experimento muitas vezes e estimar o desempenho dos classificadores nestas replicações e, por isso, normalmente adota-se $k \geq 100$. Uma das desvantagens deste método é o custo computacional.

Seguindo as práticas adotadas na literatura, o conjunto final foi dividido em 60% de treinamento e 40% de teste. As medidas de desempenho a seguir apresentadas foram calculadas com os exemplos do conjunto de teste. Para determinação dos valores dos hiperparâmetros, conceito discutido na [Subseção 1.1.1](#), o método validação cruzada foi utilizado.

5.2 Métricas de Desempenho para Problemas de Duas Classes

Matriz de Confusão

Independente de como os dados foram particionados, é necessário avaliar a qualidade dos resultados. O objetivo é comparar as respostas fornecidas pelos modelos com as respostas verdadeiras, ou seja, os rótulos originais de que se dispõe. Para problemas de classificação binários, um método adequado e bastante utilizado para analisar os resultados produzidos pelos classificadores é a matriz de confusão - e as medidas de desempenho que dela decorrem. A matriz de confusão, exibida na [Figura 22](#), é uma ideia simples, que fornece cruzadas das quantidades preditas e observadas em cada classe da variável resposta.

Na [Figura 22](#), VP são os verdadeiros positivos, ou seja, número de exemplos da classe positiva e que foram corretamente classificados; VN são os verdadeiros negativos, ou seja, o número de exemplos da classe negativa e que foram corretamente classificados; FP são os falsos positivos, ou seja, número de exemplos da classe negativa e que foram incorretamente classificados pelo modelo; e, por fim, FN são os falsos negativos, ou seja, número de exemplos da classe positiva e que foram incorretamente classificados.

Figura 22 – Matriz de confusão para um problema de classificação de duas classes.

		Classe predita	
		+	-
Classe verdadeira	+	VP	FN
	-	FP	VN

Fonte: [Faceli et al. \(2011\)](#).

As medidas de desempenho baseados na matriz de confusão e utilizadas neste trabalho são elencadas abaixo:

- Taxa de erro total (*Error Rate* - ER): a taxa de erro total nada mais é do que a razão da soma de todos os casos incorretamente classificados, independente da classe, pelo número total de objetos no conjunto de teste. Ou seja, é a soma dos elementos da diagonal secundária da matriz dividida pelo total de elementos. É formulado da seguinte maneira:

$$ER = \frac{FP + FN}{VP + VN + FN + FP}. \quad (5.1)$$

- Acurácia (*Accuracy*- ACC): a acurácia é a razão da soma de todos os casos corretamente classificados, independente da classe, pelo número total de objetos no conjunto de teste. Ou seja, é o complementar da taxa de erro e, portanto, dada por:

$$ACC = 1 - ER = \frac{VP + VN}{VP + VN + FN + FP}. \quad (5.2)$$

- Sensibilidade (*Sensitivity* - SEN): taxas altas de sensibilidade, segundo [Viana \(2015\)](#), são indicativas de modelos bons em detectar o evento de interesse, ou seja, em detectar o evento dado que o evento ocorreu. Ela corresponde, portanto, à taxa de acerto na classe positiva verdadeira e é dada por:

$$SEN = \frac{VP}{VP + FN}. \quad (5.3)$$

- Especificidade (*Specificity* - SPE): a especificidade é complementar à sensibilidade e, portanto, corresponde à taxa de acerto na classe positiva. Como afirma [Viana \(2015\)](#), ao contrário da sensibilidade, no entanto, ela indica o quão bom o modelo é para detectar não evento dado que o evento realmente não ocorreu:

$$SPE = 1 - SEN = \frac{VN}{VN + FP}. \quad (5.4)$$

- Taxa de Falso Negativo (*False Negative Rate* - FNR): também conhecida como erro do tipo I, reflete a proporção dos casos positivos incorretamente classificados e fornece, como enfatiza [Viana \(2015\)](#), a probabilidade do modelo não detectar o evento dado que o evento ocorreu. A expressão é dada por:

$$FNR = 1 - SEN = \frac{FN}{VP + FN}. \quad (5.5)$$

- Taxa de Falsos Positivos (*False Positive Rate* - FPR): conhecida como erro do tipo II, reflete a proporção de casos negativos incorretamente classificados. Estima, portanto, a probabilidade do modelo detectar o evento dado que o evento não ocorreu. É equacionada como:

$$FPR = 1 - SPE = \frac{FP}{VN + FP}. \quad (5.6)$$

Os métodos de classificação, apesar de apresentarem variável resposta categórica, produzem saídas contínuas, as quais refletem a probabilidade de um determinado indivíduo ou objeto ser classificado como um caso de sucesso (1) ou fracasso (0), ou seja, de pertencer a uma classe positiva ou negativa. Segundo [Faceli et al. \(2011\)](#), para realizar a classificação final, os valores contínuos obtidos são de alguma maneira discretizados e, para isso, emprega-se algum limiar. Ou seja, para de fato associar resultados finais binários aos indivíduos, um ponto de corte precisa ser aplicado. Um ponto de corte é um valor contínuo tal que indivíduos cuja probabilidade predita é maior ou igual ao valor estipulado serão classificados como 1 e, pelo contrário, indivíduos cuja probabilidade predita é menor do que o valor estipulado serão classificados como 0. Um ponto de corte igual a 0,5 é bastante utilizado e normalmente adotado como padrão. Pontos de corte diferentes produzirão resultados diferentes, no sentido de que a matriz de confusão será diferente e, conseqüentemente, as medidas de desempenho acima definidas. Existem diferentes maneiras de determinar o ponto de corte de ótimo e o método utilizado neste trabalho será especificado na subseção a seguir.

Análise ROC

Segundo [Fawcett \(2006\)](#), gráficos ROC são gráficos bidimensionais em que a taxa de verdadeiros positivos ([Equação 5.3](#)) é plotada no eixo das ordenadas e a taxa de falsos positivos ([Equação 5.6](#)) é plotada no eixo das abscissas. É um gráfico, portanto, que retrata o conflito de escolha entre benefícios e custos. Ainda segundo o autor, o gráfico ROC é uma técnica de visualização e seleção de classificadores baseado em seus desempenhos. A figura [Figura 23](#) exibe uma representação de um gráfico ROC.

[Faceli et al. \(2011\)](#) afirmam que a linha diagonal representa classificadores que realizam classificações aleatórias e, portanto, qualquer classificador abaixo desta linha pode ser considerado pior

Figura 23 – Gráfico ROC e área sob a curva ROC para dois classificadores distintos.

Fonte: [Fawcett \(2006\)](#).

do que o aleatório. O ponto inferior esquerdo (0,0) corresponde à estratégia de nunca produzir uma classificação positiva. Dessa maneira, um indivíduo não será classificado incorretamente como positivo, mas também não será corretamente classificado como positivo. Por outro lado, o ponto superior direito (1,1) representa classificações sempre positivas. O ponto superior esquerdo (0,1) é o ponto em que as taxas de verdadeiros positivos e verdadeiros negativos são maximizadas, ou seja, representa classificações perfeitas. Quanto mais próximo deste ponto, melhor será a classificação.

Um único resultado gera um único ponto no gráfico ROC. Uma medida também bastante utilizada é a área sob a curva ROC (*Area Under Curve* - AUC), que produz valores entre 0 e 1 e é obtida quando vários pontos são gerados e conectados. Diferentes pontos no gráfico ROC podem ser gerados empregando-se diferentes pontos de corte nas probabilidades previstas pelos métodos de classificação. O ponto (0,0) na [Figura 23](#), segundo [Habibzadeh, Habibzadeh e Yadollahie \(2016\)](#), representa também o maior ponto de corte possível e, ao avançar para a direita no gráfico, o ponto de corte diminui atingindo seu valor mínimo em (1,1). Pode-se pensar em uma curva contínua com um número infinito de pontos, mas na realidade a curva ROC é construída baseada em alguns pontos discretos. Os pontos de corte de cada um dos métodos foram obtidos por validação cruzada de modo que a distância da curva até o canto superior esquerdo do gráfico ROC fosse a menor possível, ou seja, de forma que se estivesse o mais próximo possível de uma

classificação perfeita. Os resultados dos pontos de corte de cada método serão apresentados no capítulo [Capítulo 4](#). A área sob a curva ROC foi também calculada e utilizada para comparação dos métodos de classificação, como sugerem [Faceli et al. \(2011\)](#). Para tanto, o processo de validação cruzada também foi conduzido, mas desta vez no conjunto de dados de teste. [Faceli et al. \(2011\)](#) afirmam que é comum comparar o desempenho dos algoritmos em torno de a área sob a curva ROC, e que o classificador que possuir valor mais próximo de 1 é considerado superior. Ainda segundo os autores, essa é uma maneira de deixar a análise independente do limiar empregado.

RESULTADOS

Este capítulo apresenta a especificação dos tratamentos aplicados aos dados e os resultados provenientes do emprego dos métodos de classificação abordados no [Capítulo 4](#). Os tratamentos gerais, comuns aos métodos, são exibidos na primeira seção. As seções seguintes apresentam tratamentos de dados específicos e os resultados de cada metodologia e a última seção é dedicada à compará-los.

6.1 Tratamentos Gerais dos Dados

As variáveis eliminadas de acordo com os critérios e resultados dos métodos de seleção de variáveis, apresentados na [Subseção 3.1.1](#), não foram consideradas para efeitos de classificação, exceto para o método floresta aleatória, que não necessita de nenhum tratamento diferenciado conforme o tipo da variável e consegue lidar com atributos irrelevantes ou redundantes, sem que seu desempenho seja afetado - como foi apontado na [Seção 4.4](#). Variáveis qualitativas ordinais que apresentaram alta correlação entre si, de acordo com os resultados da [Figura 13](#), também foram desconsideradas. Por isso, das variáveis que representam o *status* do pagamento, apenas a variável Atraso Máximo foi mantida.

Além disso, uma parte dos tratamentos realizados em algumas variáveis, ou mesmo no conjunto de dados completo, foi considerada adequadamente para todos os métodos de decisão. São eles:

1. **Categorização** da variável status, com o objetivo de reduzir o número de categorias deste atributo sem violar o seu conceito. Para tanto, as categorias cujos valores são maiores ou iguais a 1 foram agregadas em um único grupo, que passou a representar, portanto, clientes que atrasaram o pagamento de suas faturas por 1, 2 ou 9 meses. Esta medida é especialmente útil para o modelo de regressão logística, em que um coeficiente seria

estimado para cada uma das $c - 1$ classes da variável categórica. A distribuição da referida variável após o agrupamento é exibida na [Figura 24](#):

Figura 24 – Distribuição da variável categórica status incondicionada à variável resposta (à esquerda) e condicionada à variável resposta (à direita) após a categorização.

Fonte: Elaborado pelo autor.

2. **Normalização** das variáveis quantitativas contínuas, transformação muito utilizada e indicada na literatura para os métodos de classificação que conseguem lidar com atributos deste tipo - neste caso, KNN, Redes Neurais e Regressão Logística, não sendo necessário para floresta aleatória. A normalização é recomendável quando os limites de valores de atributos distintos são muito diferentes, para evitar que um atributo predomine sobre o outro (a menos que isso seja importante) ([FACELI et al., 2011](#)). Optou-se por proceder com a normalização por padronização, ou seja, a cada valor do atributo a ser normalizado foi subtraída uma medida de localização, a média (μ), e o valor resultante foi dividido por uma medida de escala, o desvio padrão (σ), como mostra a [Equação 6.1](#):

$$v_{norm} = \frac{v_{atual} - \mu}{\sigma}. \quad (6.1)$$

3. **Balanceamento** dos dados, para evitar que a classe majoritária da variável resposta seja favorecida na classificação. O conjunto de dados de treinamento foi balanceado de maneira

que as proporções das classes do atributo alvo fossem iguais e, por conseguinte, houve redução do número de objetos em que $y = 0$ - a classe majoritária, como mostra a [Figura 1](#). O balanceamento poderia ser feito pelo acréscimo artificial de observações da classe minoritária, também com o objetivo de igualar as proporções das classes, mas esta opção não foi considerada dada a suficiência dos dados. Vale ressaltar que não é preciso balancear o conjunto de teste, dado que o classificador é alicerçado nos objetos de treinamento. Segundo [Crone e Finlay \(2012\)](#), pesquisas sobre a amostragem de conjunto de dados desbalanceados na mineração de dados sugerem que a amostragem por redução, ou *undersampling*, produz uma precisão de classificação inferior para alguns algoritmos. Além disso, os autores concluem que para regressão logística, a técnica mais popular utilizada para construir modelos de *credit scoring* na prática, o balanceamento de dados é de menor importância. Outros métodos demonstram uma maior sensibilidade ao balanceamento, particularmente análise discriminante linear e árvores de decisão.

Além destes tratamentos, outros procedimentos foram conduzidos. A validação cruzada, metodologia descrita na [Seção 5.1](#), foi utilizada para determinação do melhor ponto de corte, como já foi mencionado, e dos valores de hiperparâmetros, conceito retratado na [Subseção 1.1.1](#). O método eleito para ajustar os hiperparâmetros foi o *grid search*, que consiste em construir um modelo para todos os valores dos hiperparâmetros dentro de um intervalo estabelecido, avaliar cada um deles e selecionar aquele que produz o melhor resultado. Neste caso, o valor escolhido, ou a combinação de valores, foi aquele que resultou em uma menor taxa de erro média.

6.2 Desempenho dos Classificadores

Classificador Naive Bayes

O classificador Naive Bayes exige que sejam calculadas a probabilidade *a priori* $P(y_c)$ de observar uma classe y_c e a probabilidade condicional de observar um valor de uma variável explicativa dado que o objeto pertence uma determinada classe. Para o primeiro caso, como sugerem [Faceli et al. \(2011\)](#), basta manter um contador para cada classe. Para o segundo caso, no entanto, é preciso fazer distinção entre atributos qualitativos e quantitativos. Para atributos qualitativos, segundo [Yang e Webb \(2003\)](#), suas probabilidades podem ser estimadas a partir das frequências correspondentes. Para um atributo quantitativo, por sua vez, a estimativa da densidade de probabilidade ou a discretização podem ser empregadas para estimar suas probabilidades. A primeira abordagem requer suposição sobre a forma da distribuição de probabilidades a partir da qual os valores foram extraídos. Se a densidade presumida não é uma estimativa adequada da densidade real, como enfatizam os autores, o desempenho do método de classificação tende a se degradar. Como a densidade real é normalmente desconhecida, a discretização é uma alternativa viável. Discretizar uma variável x^j é associar cada um de seus valores a um intervalo numérico $[a, b)$, com $a < b$, e utilizar como novo valor a categoria do intervalo correspondente.

Há diferentes maneiras de criar intervalos aos quais os valores das variáveis serão associados, e neste trabalho optou-se por adotar os valores mínimo e máximo e os quartis. Portanto, para cada variável contínua, quatro intervalos foram criados.

A validação cruzada para determinação do ponto de corte gerou os resultados apresentados na Tabela 12 e a curva ROC exibida na Figura 25:

Tabela 12 – Ponto de corte baseado na especificidade e sensibilidade para cada amostra k da validação cruzada.

Ponto de Corte - <i>Naive Bayes</i>											
Medida	1	2	3	4	5	6	7	8	9	10	Média
Sensitividade	0,68	0,70	0,61	0,65	0,73	0,65	0,73	0,71	0,63	0,71	0,55
Especificidade	0,67	0,66	0,69	0,72	0,65	0,73	0,66	0,64	0,64	0,67	
Ponto de Corte	0,59	0,51	0,62	0,61	0,49	0,64	0,46	0,52	0,56	0,54	

Figura 25 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para determinação do ponto de corte para o método Naive Bayes.

Fonte: Elaborado pelo autor.

Após a determinação do ponto do corte, a taxa de erro do modelo e as outras medidas de desempenho, descritas na Seção 5.2, foram geradas a partir do conjunto de exemplos de teste. Os resultados são exibidos na Tabela 13:

Tabela 13 – Medidas de desempenho do método Naive Bayes obtidas a partir do conjunto de dados de teste.

Medidas de Desempenho - Naive Bayes					
ER	ACC	SEN	SPE	FNR	FPR
0,346	0,654	0,655	0,654	0,345	0,346

A comparação dos resultados revela que a taxa de erro alcançada é 64,6% superior à taxa obtida por Yeh e Lien (2009), igual a 0,21. Esta diferença pode ter sido ocasionada pelos tratamentos realizados e pelas variáveis eliminadas na etapa de seleção de atributos. Variáveis redundantes, por exemplo, podem melhorar a acurácia dos modelos - e, consequentemente, reduzir o erro de classificação.

O pacote de dados chamado "e1071", do repositório "CRAN", foi utilizado para compilação do método.

K-Nearest Neighbor

Dado que o *k*-vizinho mais próximo é baseado em distâncias, atributos categóricos precisam ser convertidos em números. Se o atributo é qualitativo nominal com apenas duas categorias, então codificá-lo para 0 – 1 é razoável. Se, por outro lado, o número de categorias é maior do 2, variáveis *dummy* podem ser criadas de forma que a distância entre quaisquer pares de valores seja sempre a mesma. Isto pode ocorrer, por exemplo, quando se quer calcular a distância entre dois objetos diferentes da base considerando uma variável categórica com 3 níveis, que podem ser "Vermelho", "Verde" ou "Amarelo". Como o atributo qualitativo é nominal, ou seja, a ordem das classes não é determinante, não se pode atribuir a cada categoria um número inteiro, pois a distância calculada pode refletir diferenças entre categorias. Ao criar uma variável *dummy* para cada classe do atributo nominal, garante-se que diferenças como estas não ocorrerão. Se o atributo categórico é ordinal, números inteiros podem ser associadas às categorias de maneira que a ordenação entre elas seja preservada. Neste caso, o que se quer é observar diferenças nas distâncias calculadas sobre uma categoria ou outra. Se as variáveis preditoras são quantitativas contínuas ou discretas, as distâncias podem ser facilmente calculadas.

O número *k* de vizinhos, hiperparâmetro do modelo, foi estipulado também por validação cruzada, como acima mencionado. Foram testados números ímpares de 1 a 51, e o que resultou no menor erro médio foi $k = 19$, como mostra a figura Figura 26:

Estabelecido o número de vizinhos, procedeu-se com a validação cruzada para determinação do ponto de corte, que gerou os resultados apresentados na Tabela 14 e a curva ROC exibida na Figura 27:

Os resultados das medidas de desempenho do modelo são exibidos na Tabela 15:

A taxa de erro obtida por Yeh e Lien (2009) é igual a 0,160, 49% menor do que a taxa exibida na Tabela 15.

Figura 26 – Taxa de erro média obtida através de validação cruzada com $k = 10$ para determinação do valor do hiperparâmetro número de vizinhos.

Fonte: Elaborado pelo autor.

Tabela 14 – Ponto de corte baseado na especificidade e sensibilidade para cada amostra k da validação cruzada.

Ponto de Corte - KNN											
Medida	1	2	3	4	5	6	7	8	9	10	Média
Sensitividade	0,66	0,65	0,70	0,66	0,70	0,70	0,68	0,71	0,63	0,67	0,51
Especificidade	0,69	0,75	0,69	0,75	0,75	0,72	0,74	0,72	0,69	0,72	
Ponto de Corte	0,53	0,53	0,53	0,50	0,53	0,48	0,53	0,53	0,53	0,48	

O pacote de dados chamado "class", do repositório "CRAN", foi utilizado para compilação do método.

Regressão Logística

Em regressão logística (ou regressão linear), há também um problema específico relacionado à atributos qualitativos com três ou mais categorias. Segundo [Miles e Shevlin \(2001\)](#), ainda que as categorias sejam numéricas, a variável é tratada pelo modelo de regressão como uma variável contínua, o que produz resultados inconsistentes. Existem algumas maneiras de codificar dados categóricos, e duas das mais conhecidas são *dummy coding* e *effect coding*. Na primeira

Figura 27 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para determinação do ponto de corte para o método KNN.

Fonte: Elaborado pelo autor.

Tabela 15 – Medidas de desempenho do método KNN obtidas a partir do conjunto de dados de teste.

Medidas de Desempenho - KNN					
ER	ACC	SEN	SPE	FNR	FPR
0,314	0,686	0,655	0,695	0,345	0,305

abordagem, um dos grupos é considerado o grupo de referência, interpretado como a condição de controle, e variáveis *dummy* são criadas para identificar em quais condições os outros sujeitos estão - na prática, $c - 1$ *dummies* são criadas. Nem sempre o grupo de referência é óbvio e, nestes casos, a segunda abordagem é recomendada. A principal diferença entre as abordagens é que, com *dummy coding*, um grupo é usado como referência e tem seus valores transformados para 0, mas com *effect coding* um grupo é arbitrariamente selecionado e à ele é atribuído valor igual a -1 . Neste trabalho, a abordagem *effect coding* foi adotada. Além deste tratamento, é importante destacar que o modelo final foi selecionado por meio da seleção *backward*.

A validação cruzada para determinação do ponto de corte gerou os resultados apresentados na Tabela 16 e a curva ROC exibida na Figura 28:

Os resultados das medidas de desempenho do modelo são exibidos na Tabela 17:

Tabela 16 – Ponto de corte baseado na especificidade e sensibilidade para cada amostra k da validação cruzada.

Ponto de Corte - Regressão Logística											
Medida	1	2	3	4	5	6	7	8	9	10	Média
Sensibilidade	0,65	0,68	0,63	0,67	0,73	0,68	0,68	0,68	0,66	0,65	0,49
Especificidade	0,74	0,71	0,76	0,80	0,67	0,74	0,75	0,72	0,75	0,74	
Ponto de Corte	0,48	0,46	0,53	0,63	0,41	0,49	0,44	0,44	0,48	0,53	

Figura 28 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para determinação do ponto de corte para o método Regressão Logística.

Fonte: Elaborado pelo autor.

A taxa de erro obtida por [Yeh e Lien \(2009\)](#) para este método é igual a 0,180, menor do que a taxa reportada para o método KNN. As tabelas [Tabela 15](#) e [Tabela 17](#) mostram, pelo contrário, que a taxa de erro da regressão logística é 14% menor do que a taxa do método KNN. Além disso, fazendo uma comparação direta das metodologias, a taxa de erro do artigo é 33% menor.

A função "glm" do *software* R foi utilizada para compilação do método.

Tabela 17 – Medidas de desempenho do método regressão logística obtidas a partir do conjunto de dados de teste.

Medidas de Desempenho - Regressão Logística					
ER	ACC	SEN	SPE	FNR	FPR
0,269	0,731	0,665	0,749	0,335	0,251

Random Forests

Para a metodologia *Random Forests*, nenhum tratamento adicional foi aplicado. Pelo contrário, todos os atributos foram consideradas com o intuito de comparar o seu poder preditivo com o dos outros métodos, os quais de partida englobaram menos variáveis explicativas.

O número de árvores na floresta aleatória, um dos hiperparâmetros do modelo, foi escolhido a partir do erro OOB. O número foi selecionado que propiciou o melhor custo benefício, ou seja, não tão custoso computacionalmente e cujo erro OOB associado é apenas ligeiramente superior ao produzido por florestas com mais árvores. A [Figura 29](#) exhibe o erro OOB de acordo com o número de árvores na floresta, a uma linha foi traçada no número de árvores escolhido.

Figura 29 – Taxa de erro OOB de acordo com o número de árvores na floresta aleatória.

Fonte: Elaborado pelo autor.

O número de variáveis selecionadas a cada quebra, hiperparâmetro chamado de *mtry*, foi estipulado por validação cruzada. Foram testados números de 1 a 15, e o que resultou no segundo menor erro OOB foi *mtry*= 4, como mostra a figura [Figura 30](#):

Figura 30 – Taxa de erro OOB média obtida através de validação cruzada com $k = 10$ para determinação do valor do hiperparâmetro *mtry*.

Fonte: Elaborado pelo autor.

A validação cruzada para determinação do ponto de corte gerou os resultados apresentados na [Tabela 18](#) e a curva ROC exibida na [Figura 31](#):

Tabela 18 – Ponto de corte baseado na especificidade e sensibilidade para cada amostra k da validação cruzada.

Ponto de Corte - <i>Random Forests</i>											
Medida	1	2	3	4	5	6	7	8	9	10	Média
Sensibilidade	0,75	0,69	0,67	0,71	0,71	0,66	0,71	0,68	0,60	0,70	0,50
Especificidade	0,68	0,73	0,74	0,74	0,80	0,77	0,76	0,78	0,77	0,73	
Ponto de Corte	0,46	0,47	0,50	0,47	0,52	0,52	0,48	0,54	0,52	0,47	

Os resultados das medidas de desempenho do modelo são exibidos na [Tabela 19](#):

[Yeh e Lien \(2009\)](#) reportaram uma taxa de erro de validação igual a 0,17 para para árvores de decisão, 41% menor do que o resultado exibido na [Tabela 19](#).

Figura 31 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para determinação do ponto de corte para o método *Random Forests*.

Fonte: Elaborado pelo autor.

Tabela 19 – Medidas de desempenho do método *random forests* obtidas a partir do conjunto de dados de teste.

Medidas de Desempenho - <i>Random Forests</i>					
ER	ACC	SEN	SPE	FNR	FPR
0,289	0,711	0,644	0,730	0,356	0,270

O pacote "randomForest" do *software* R foi utilizado para compilação do método.

Redes Neurais

Em redes neurais, o tratamento aplicado aos atributos nominais foi também o *effect coding*, o mesmo utilizado em regressão logística. Para inicializar os pesos da rede tomou-se valores no intervalo $\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right]$, em que n é o número de nós na camada oculta à que os pesos estão associados. Segundo Marsland (2014), se os pesos são grandes, então a ativação de um neurônio provavelmente ficará próxima de 0 ou 1, o que significa que os gradientes são pequenos e, portanto, o aprendizado é muito lento. O número máximo de iterações estipulado foi $1e + 05$ e o algoritmo escolhido foi *resilient backpropagation*.

O número de camadas e o número de neurônios em cada camada, hiperparâmetros do modelo, foram determinados também por validação cruzada. Dois grupos de testes foram conduzidos: um deles considerou 1 camada oculta e 1 a 10 neurônios; e o segundo considerou 2 camadas ocultas e 1 a 5 neurônios em cada camada. No total, foram testadas 35 possibilidades, sendo 10 referentes ao primeiro grupo de testes e 25 ao segundo. Similarmente aos outros métodos, os valores escolhidos para os hiperparâmetros foram aqueles que geraram a menor taxa de erro. A [Figura 32](#) exibe as taxas de erro obtidas com o primeiro grupo de testes e a [Figura 33](#) as taxas obtidas com o segundo grupo.

Figura 32 – Taxa de erro média obtida através de validação cruzada com $k = 10$ para determinação do número de camadas ocultas e do número de neurônios em cada camada.

Fonte: Elaborado pelo autor.

Determinados o número de camadas e o número de neurônios em cada camada, procedeu-se com validação para determinação do ponto de corte, que gerou os resultados apresentados na [Tabela 20](#) e a curva ROC exibida na [Figura 34](#):

Os resultados das medidas de desempenho do modelo são exibidos na [Tabela 21](#):

A taxa de erro de validação obtida por [Yeh e Lien \(2009\)](#) para este método é igual a 0,170, igual à taxa reportada para o método árvores de decisão e menor apenas do que a taxa reportada para o método KNN. Fazendo uma comparação direta das metodologias, a taxa de erro

Figura 33 – Taxa de erro média obtida através de validação cruzada com $k = 10$ para determinação do número de camadas ocultas e do número de neurônios em cada camada.

Fonte: Elaborado pelo autor.

Tabela 20 – Ponto de corte baseado na especificidade e sensibilidade para cada amostra k da validação cruzada.

Ponto de Corte - Redes Neurais											
Medida	1	2	3	4	5	6	7	8	9	10	Média
Sensitividade	0,64	0,68	0,66	0,67	0,71	0,72	0,68	0,69	0,68	0,68	0,47
Especificidade	0,77	0,77	0,75	0,77	0,76	0,74	0,76	0,76	0,68	0,73	
Ponto de Corte	0,54	0,46	0,50	0,45	0,46	0,43	0,47	0,50	0,39	0,471	

do artigo é 35% menor.

A função "neuralnet" do *software* R foi utilizada para compilação do método.

Figura 34 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para determinação do ponto de corte para o método Redes Neurais.

Fonte: Elaborado pelo autor.

Tabela 21 – Medidas de desempenho do método redes neurais obtidas a partir do conjunto de dados de teste.

Medidas de Desempenho - Redes Neurais					
ER	ACC	SEN	SPE	FNR	FPR
0,264	0,736	0,635	0,765	0,365	0,235

6.3 Comparativo de Desempenho

A [Figura 35](#) e a [Tabela 22](#) exibem, respectivamente, as curvas ROC e as medidas de desempenho calculadas para cada um dos métodos de classificação. Como as curvas ROC foram obtidas utilizando-se validação cruzada com $k = 10$, a curva exibida para cada um dos métodos corresponde à média da validação cruzada. Considerando os valores da área sob a curva ROC (AUC), é possível observar que o método *random forests* produziu o maior resultado, seguido pelos métodos regressão logística e redes neurais. Um ponto importante é que o método *random forests* é o único que contemplou todas as variáveis explicativas. Isto significa que, mesmo apesar de o custo computacional aumentar à medida que aumenta a dimensão dos dados, nenhum tratamento específico com relação ao tipo da variável é necessário. Por sua vez, os métodos

redes neurais e regressão logística apresentaram taxas de acurácia total (ACC) maiores do que as taxas de acurácia dos outros métodos. O método Naive Bayes foi o que apresentou a menor taxa de acurácia total. A taxa de falsos negativos (FNR) mostra que os métodos KNN e Naive Bayes cometem com menos frequência o erro do tipo I quando comparados aos métodos *random forests* e redes neurais. Além disso, os métodos os métodos KNN e Naive Bayes exibem taxas de sensibilidade menores somente do que a taxa de sensibilidade do método regressão logística. O método redes neurais, por outro lado, comete com menos frequência o erro do tipo II, o qual está relacionado a classificar indevidamente um exemplo negativo como positivo. A taxa de sensibilidade deste método é a menor, indicando que ele é menos eficiente em detectar o evento de interesse.

Tabela 22 – Medidas de desempenho dos métodos de classificação.

Método de Classificação	ER	SEN	SPE	FNR	FPR	AUC Validação Cruzada
Naive Bayes	0,346	0,655	0,654	0,345	0,346	0,712
KNN	0,314	0,655	0,695	0,345	0,305	0,700
Regressão Logística	0,269	0,665	0,749	0,335	0,251	0,737
Random Forests	0,289	0,644	0,730	0,356	0,270	0,745
Redes Neurais	0,264	0,635	0,765	0,365	0,235	0,731

Figura 35 – Curvas ROC produzidas a partir de validação cruzada com $k = 10$ para comparação dos modelos de classificação.

Fonte: Elaborado pelo autor.

CONCLUSÃO

Os métodos de aprendizado de máquina estudados podem ser analisados sob diferentes aspectos. O método *random forest* é melhor quando se considera o valor da área sob a curva ROC (AUC), medida que não está atrelada aos pontos de corte escolhidos para a classificação. Uma das vantagens do método *random forests*, além de produzir bons resultados, é a facilidade em lidar com atributos numéricos ou categóricos, e por isso também é bastante utilizado em procedimentos de seleção de variáveis. Por envolver muitas árvores de decisão, no entanto, a sua interpretabilidade é reduzida. A análise da taxa de erro total (ER), por sua vez, revela que o método redes neurais comete menos classificações incorretas em comparação aos outros métodos. Apesar de menor, a taxa de erro total do método regressão logística é bem próxima à taxa de erro total do método redes neurais. Por levar em consideração os erros cometidos nas duas classes, no entanto, a taxa de erro total pode não ser a medida de desempenho mais apropriada. Em análise de risco ou *credit scoring*, o custo de atribuir o rótulo de bom pagador a um mau pagador é mais alto do que atribuir o rótulo de mau pagador a um bom pagador, e a medida que reflete este tipo de erro é a taxa de falsos negativos (FNR). Redes neurais é o método que apresenta a maior taxa de falsos negativos. Isto significa que este método classifica clientes maus pagadores como bons pagadores com mais frequência do que os outros métodos avaliados. O método regressão logística comete com menos frequência um erro deste tipo. Os métodos KNN e Naive Bayes exibem taxas de erro total maiores e valores da medida AUC menores, mas taxas de falsos negativos melhores do que os métodos redes neurais e *random forest*. A taxa de sensibilidade revela também que o método regressão logística é o melhor em classificar um indivíduo como evento dado que ele realmente é evento. Pelo contrário, o método redes neurais é o melhor em classificar um indivíduo como não evento dado que ele realmente não é evento.

Após as análises conduzidas neste trabalho, o método regressão logística é considerado o que possui a melhor relação custo-benefício. Isto porque este método apresenta um bom preditivo, comete com menos frequência um erro do tipo I e é mais facilmente interpretável.

REFERÊNCIAS

- AGGARWAL, C. C. **Data Classification: Algorithms and Applications**. 1st. ed. [S.l.]: Chapman & Hall/CRC, 2014. Citado nas páginas 48 e 49.
- AGRESTI, A. **Categorical Data Analysis**. 2nd. ed. [S.l.]: Wiley, 2003. 734 p. Citado nas páginas 50 e 52.
- BERRY, K. J.; JOHNSTON, J. E.; JR., P. W. M. **A Chronicle of Permutation Statistical Methods: 1920–2000, and beyond**. 1st. ed. [S.l.]: Springer, 2014. 517 p. Citado nas páginas 41, 53 e 55.
- BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. 4th. ed. [S.l.]: Atual, 1987. 321 p. Citado nas páginas 28, 30, 39 e 50.
- CHOK, N. S. **Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data**. Dissertação (Master Thesis) — University of Pittsburgh, 2010. Citado nas páginas 39 e 42.
- CHOU, M. **Cash and credit card crisis in Taiwan**. 2006. 24-27 p. Citado na página 22.
- CRONE, S.; FINLAY, S. Instance sampling in credit scoring: An empirical study of sample size and balancing. **International Journal of Forecasting**, v. 28, 03 2012. Disponível em: <[http://www.sven-crone.com/papers/Crone,%20Finlay%20\(2011\)%20Instance%20Sampling%20in%20credit%20Scoring%20V6.pdf](http://www.sven-crone.com/papers/Crone,%20Finlay%20(2011)%20Instance%20Sampling%20in%20credit%20Scoring%20V6.pdf)>. Citado na página 91.
- CYBENKO, G. Approximation by superpositions of a sigmoidal function. **Mathematics of Control, Signals, and Systems**, v. 2, p. 303–314, 1989. Citado na página 76.
- DOMINGOS, P.; PAZZANI, M. On the optimality of the simple bayesian classifier under zero-one loss. **Mach. Learn.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 29, n. 2-3, p. 103–130, nov. 1997. ISSN 0885-6125. Disponível em: <<https://doi.org/10.1023/A:1007413511361>>. Citado na página 59.
- FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. **Inteligência Artificial: Uma abordagem de Aprendizado de Máquina**. 1st. ed. [S.l.]: LTC, 2011. ISBN 8521618808, 9788521618805. Citado nas páginas 18, 19, 22, 27, 28, 29, 38, 47, 48, 58, 59, 60, 66, 67, 68, 69, 73, 76, 77, 78, 79, 81, 83, 84, 85, 86, 88, 90 e 91.
- FAWCETT, T. An introduction to roc analysis. v. 27, n. 8, p. 861–874, 06 2006. Citado nas páginas 86 e 87.
- GU, Q.; LI, Z.; HAN, J. Generalized fisher score for feature selection. In: **Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence**. AUAI Press, 2011. p. 266–273. Disponível em: <<http://dl.acm.org/citation.cfm?id=3020548.3020580>>. Citado nas páginas 48 e 53.
- GUNTHER, F.; FRITSCH, S. neuralnet: Training of neural networks. **The R Journal**, ISSN, v. 2, n. 1, p. 30–38, 2010. Citado nas páginas 73, 77 e 80.

- HABIBZADEH, F.; HABIBZADEH, P.; YADOLLAHIE, M. On determining the most appropriate test cut-off value: The case of tests with continuous results. v. 26, p. 297–307, 10 2016. Citado na página 87.
- HAUKE, J.; KOSSOWSKI, T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. v. 30, n. 2, p. 87–93, 04 2011. Citado na página 41.
- HAYKIN, S. **Redes Neurais: Princípios e Prática**. 2nd. ed. [S.l.]: Bookman, 2001. ISBN 0132733501, 9780132733502. Citado nas páginas 74, 75 e 76.
- HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. 2nd. ed. [S.l.]: Wiley, 2000. ISBN 0471356328, 9780471356325. Citado nas páginas 61, 64 e 65.
- JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with Applications in R**. [S.l.: s.n.], 2013. Springer. Citado nas páginas 19, 60, 62, 64, 65, 66, 67, 69 e 70.
- KHAMIS, H. Measures of association: How to choose? **Journal of Diagnostic Medical Sonography**, v. 24, n. 3, p. 155–162, 2008. Disponível em: <<https://doi.org/10.1177/8756479308317006>>. Citado nas páginas 51, 53 e 55.
- KONONENKO, I. Semi-naive bayesian classifier. In: **Proceedings of the European Working Session on Learning on Machine Learning**. Berlin, Heidelberg: Springer-Verlag, 1991. (EWSL-91), p. 206–219. ISBN 0-387-53816-X. Disponível em: <<http://dl.acm.org/citation.cfm?id=112152.112167>>. Citado na página 59.
- LIEBETRAU, A. M. **Measures of Association**. [S.l.]: SAGE, 1983. v. 32. 95 p. Citado na página 51.
- LIRA, S. A.; NETO, A. C. Coeficientes de correlação para variáveis ordinais dicotômicas derivados do coeficiente linear de pearson. **RECIE**, v. 15, n. 1/2, p. 45–53, 2006. Citado na página 55.
- LUXBURG, U. von; SCHÖLKOPF, B. Statistical learning theory: Models, concepts, and results. In: . [S.l.: s.n.], 2008. Citado nas páginas 17 e 18.
- MARSLAND, S. **Machine Learning: An Algorithmic Perspective, Second Edition**. 2nd. ed. [S.l.]: Chapman & Hall/CRC, 2014. ISBN 1466583282, 9781466583283. Citado nas páginas 17, 18, 19, 20, 59, 71, 73, 75, 76, 78, 79 e 99.
- MILES, J.; SHEVLIN, M. **Applying Regression and Correlation: A Guide for Students and Researchers**. 1st. ed. [S.l.]: SAGE, 2001. Citado na página 94.
- MITCHELL, T. M. **Machine Learning**. [S.l.: s.n.], 1997. McGraw-Hill. Citado nas páginas 17, 22, 57, 58, 59, 71, 75, 79 e 84.
- VIANA, R. F. **Técnicas de classificação aplicadas a credit scoring: revisão sistemática e comparação**. Dissertação (Dissertação) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2015. Citado nas páginas 84, 85 e 86.
- WANG, Z.; SUDA, R.; EGAWA, R.; TAKIZAWA, H. **Auto-tuning of Hyperparameters of Machine Learning Models**. 2018. HPC Asia. Citado na página 20.

YANG, Y.; WEBB, G. I. On why discretization works for naive bayesian classifiers. **AI 2003: Advances in Artificial Intelligence**, Springer-Verlag, v. 2903, p. 440–452, 2003. Disponível em: <<http://web.cs.iastate.edu/~jtian/cs573/Papers/Yang-bayes-03.pdf>>. Citado na página 91.

YEH, I.-C.; LIEN, C.-h. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. **Expert Syst. Appl.**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 36, n. 2, p. 2473–2480, mar. 2009. ISSN 0957-4174. Disponível em: <<http://dx.doi.org/10.1016/j.eswa.2007.12.020>>. Citado nas páginas 22, 93, 96, 98 e 100.

ZHAO, Z.; MORSTATTER, F.; SHARMA, S.; ALELYANI, S.; ANAND, A.; LIU, H. Advancing feature selection research. p. 1–28, 01 2010. Citado na página 48.