Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

MÁQUINAS DE VETORES SUPORTE COM APLICAÇÃO EM CLASSIFICAÇÃO DE CRÉDITO

Bruno Matheus Brandini

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

MÁQUINAS DE VETORES SUPORTE COM APLICAÇÃO EM CLASSIFICAÇÃO DE CRÉDITO

Bruno Matheus Brandini
Orientador: Prof. Dr. Ricardo Felipe Ferreira
Coorientadora: Profa. Dra. Daiane Aparecida Zuanetti

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs-UFSCar, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Março de 2023

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

SUPPORT VECTOR MACHINE APPLIED TO CREDIT RISK

Bruno Matheus Brandini
Advisor: Prof. Dr. Ricardo Felipe Ferreira
Co-advisor: Profa. Dra. Daiane Aparecida Zuanetti

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
March 2023

Bruno Matheus Brandini

MÁQUINAS DE VETORES SUPORTE COM APLICAÇÃO EM CLASSIFICAÇÃO DE CRÉDITO

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Bruno Matheus Brandini e aprovado pela banca examinadora.

Aprovado em 22 de Março de 2023.

Banca Examinadora:

- Prof. Dr. Ricardo Felipe Ferreira
- Prof. Dr. Luis Ernesto Bueno Salasar
- Prof. Dr. Luis Aparecido Milan

Dedico este trabalho aos meus pais que sempre me apoiaram e me deram a base para chegar até aqui.

Agradecimentos

Ao finalizar esta etapa da minha vida, vejo que o trabalho aqui apresentado possui muitos autores que contribuíram para seu desenvolvimento, desde o suporte técnico, teórico e emocional. Todos foram essenciais para realização desse projeto e com toda minha admiração, quero expressar meus sinceros agradecimentos:

A Deus, pela orientação de cada dia, por ter me fortalecido em todos os momentos que achei que não conseguiria.

Aos meus pais, Aparecido e Cristiana, minha profunda admiração e reconhecimento por todos os sacrifícios que fizeram para que eu pudesse conquistar meus objetivos e por sempre acreditarem em mim. Ao meu irmão Gabriel, agradeço pelo companheirismo e conversas que foram importantes distrações nos momentos de apreensão.

A minha namorada Emanuelle por todas as palavras de incentivo, pelo companheirismo e pela compreensão que teve comigo durante esse período, seu apoio foi fundamental para a conclusão desse trabalho.

Ao meu orientador Ricardo Ferreira e a minha coorientadora Daiane Zuanetti, pela orientação e amizade durante todo esse período. Sou grato por todos os ensinamentos, conselhos e atendimentos, principalmente aqueles ao final das aulas. Obrigado por terem acreditado em meu potencial e pelas oportunidades que me foram abertas nesse período. Vocês tem todo meu respeito e admiração.

Por fim, reconheço também todas as pessoas que de alguma forma contribuíram para que eu pudesse me tornar a pessoa e o profissional que sou hoje.

"The greater our knowledge increases the more our ignorance unfolds."

(John F. Kennedy)

Resumo

A concessão de crédito representa um dos produtos com maior rentabilidade dentro de uma instituição financeira. Entretanto, para garantir lucro é primordial que as instituições saibam a quem estão concedendo seu capital. Nesse cenário, uma ferramenta fundamental para auxiliar na tomada de decisão da concessão de recursos é a classificação de crédito, que tem a finalidade de predizer a qual classe um cliente pertence, se ele tem um comportamento inadimplente ou adimplente. Logo, é de suma importância que essa ferramenta reproduza resultados próximos da realidade, com baixa margem de erro para assim evitar prejuízos financeiros para a empresa concedente do crédito. Entretanto, na conjuntura de análise de crédito, os bancos de dados são, em sua maioria, desbalanceados, uma vez que contém mais observações referentes a clientes adimplentes (classe majoritária) do que clientes inadimplentes (classe minoritária), que pode acarretar em um viés de classificação. Como alternativa para superar tal viés na classificação, podemos aplicar um pré-processamento no conjunto de dados, visando equilibrar as classes, ou realizar modificações no algoritmo de classificação, para que este possa lidar adequadamente com o problema de desequilíbrio de classe. Portanto, esse trabalho tem como proposta aplicar o classificador de máquina de vetores suporte, no contexto de classificação de crédito, para discriminação de clientes solicitantes de crédito, comparando o desempenho da técnica tanto em conjuntos de dados balanceados, a partir do método de sobreamostragem SMOTE, como também em dados desbalanceados, ao qual também aplicaremos a metodologia de máquina de vetores suporte sensível ao custo, uma técnica proposta para lidar com o desequilíbrio de classes. Além disso, compararemos o desempenho do classificador de máquina de vetores suporte com outros classificadores habitualmente utilizados no cenário de crédito, como a regressão logística e floresta aleatória. Para esse fim, o estudo será aplicado em dados reais, e avaliado em termos de algumas métricas que mensuram o desempenho de predição.

Palavras-chave: *Classificação, Crédito, Dados desbalanceados, Máquinas de vetores suporte, SMOTE.*

Abstract

The credit granting represents one of the products with the highest profitability within a financial institution. However, to ensure profit, institutions must know to whom they lend their capital. In this scenario, a fundamental tool to assist in decision-making regarding the granting of funds is the credit risk which purpose is to predict the creditworthiness of a borrower, classifying the customer as non-defaulting or a defaulting customer. Therefore, this tool must reproduce results close to reality with a low margin of error to avoid financial losses for the credit-granting institution. Nonetheless, in the context of credit analysis, the databases used in the credit risk contain more observations referring to non-defaulting customers (majority class) than defaulting customers (minority class) turning them imbalanced and prone to lead to bias in credit risk. Alternatives to overcome such bias in the classification and adequately deal with the problem of class imbalance is to apply a pre-processing in the data set to balance the classes or modify the classification algorithm. Therefore, in the credit risk context, this work proposes to apply the support vector machine classifier in the discrimination of customers requesting a loan, comparing the performance of this technique both in balanced and imbalanced data sets. In the former will be used the oversampling SMOTE method and in the later the cost-sensitive support vector machine methodology since it is proposed to deal with imbalaced classes. Furthermore, this work compare the performance of the support vector machine classifier with other classifiers commonly used in the credit scenario, such as logistic regression and random forest. The study will be applied to real data and evaluated regards to some metrics that measure the prediction performance.

Keywords: *Classification, Credit, Imbalanced data, Support vector machine, SMOTE.*

Lista de Figuras

3.1 Hiperplano gerado pela equação $0,2 + 0,85x_1 - 1x_2 = 0$, representado pela reta em preto. Os pontos em azul representam o conjunto de pares ordenados que satisfazem $0,2 + 0,85x_1 - 1x_2 < 0$, e os pontos em vermelho representam o conjunto de pares ordenados que satisfazem $0,2 + 0,85x_1 - 1x_2 > 0$. 42

3.2 Hiperplanos $1 + x_1 - x_2 = 0$, $1 + 2x_1 - 0,5x_2 = 0$ e $1 + 2x_1 - x_2 = 0$ representados pelas retas em preto. Os pontos em azul representam o conjunto de pares ordenados associados aos clientes adimplentes e os pontos em vermelho representam o conjunto de pares ordenados associados aos clientes inadimplentes. 43

3.3 Hiperplano $0,8 + 1,2x_1 - x_2 = 0$ representado pela reta em preto. Os pontos em azul representam o conjunto de pares ordenados associados aos clientes adimplentes e os pontos em vermelho representam o conjunto de pares ordenados associados aos clientes inadimplentes. 44

3.4 Os pontos em azul são os pares ordenados associados aos clientes adimplentes e os pontos em vermelho são os pares ordenados associados aos clientes inadimplentes. O hiperplano de margem máxima está representado pela linha em preto. Para este exemplo, a margem está representada pela linha pontilhada, que simboliza a menor distância de uma observação ao hiperplano. Os pontos azuis e vermelhos que estão sobre a linhas pontilhadas são os vetores suporte. 46

3.5 Os pontos em azul representam os clientes adimplentes e os pontos em vermelho os clientes inadimplentes. Nesse caso, as duas classes não são linearmente separáveis, ou seja, não conseguimos encontrar um hiperplano $H_{\boldsymbol{\beta}}$, tal que as duas classes sejam separadas perfeitamente. Dessa forma, não é possível aplicar o classificador de margem máxima. 50

3.6 Os pontos em azul simbolizam os clientes adimplentes e os pontos em vermelho os inadimplentes. As observações representadas por um "quadrado"são os vetores suporte que orientam a obtenção do hiperplano. As observações 5, 10 e 12 estão violando o limite de sua respectiva margem, logo são observações classificadas erroneamente, de modo que $\epsilon_5, \epsilon_{10}, \epsilon_{12}$ são valores entre 0 e 1 . Para a observação 3, $\epsilon_3 > 1$ dado que está classificada do lado errado do hiperplano. 52

3.7 No plano cartesiano à esquerda temos a representação de um conjunto de treinamento bidimensional não linear. Após a aplicação da função Φ (futuramente definiremos como *Kernel*) obtemos um novo conjunto de treinamento com dimensionalidade três, representado pelo plano cartesiano à direita, no qual torna-se linear no espaço de características. **Fonte:** Hachimi *et al.* (2020). 53

3.8 Representação do hiperplano ajustado em um cenário de desequilíbrio de classes. Os pontos representados por " - "em azul simbolizam a classe majoritária (clientes adimplentes) enquanto que os pontos representados por " + " simbolizam a classe minoritária (clientes inadimplentes). A linha em preto pontilhada simboliza o limite ideal para separação e a linha em preto preenchida representa o hiperplano obtido com a aplicação do classificador de máquina de vetores suporte. **Fonte:** Adaptado de Phoungphol *et al.* (2012) 60

4.1 Correlograma entre as variáveis utilizadas no estudo, em que cada quadrado indica o valor e a intensidade da correlação entre as referidas variáveis. A intensidade é ilustrada a partir das cores dos quadrados, de modo que tons mais fortes de vermelho expressam forte correlação negativa, tons mais claros e próximos de branco expressam baixa correlação, sendo o branco a indicação de independência entre as variáveis, por fim tons mais escuros de azul expressam uma correlação forte e positiva. A simbologia V1,. . . ,V6, representa de forma simplificada as seis variáveis selecionadas para prosseguimento das análises, ao passo que: V1 = Idade, V2 = Atraso máximo, V3 = Proporção do limite utilizado, V4 = Proporção limite x pagamento, V5 = Proporção de pagamento e V6 = Quantidade de atraso. 72

4.2 Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para Idade. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Atraso máximo. . 73

4.3 Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para a variável Proporção do limite utilizado. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Proporção do pagamento em relação ao limite. 74

4.4 Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para a variável Proporção de pagamento. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Quantidade de atrasos. 75

4.5 Correlograma entre as variáveis utilizadas no estudo, em que cada quadrado indica o valor da correlação entre as referidas variáveis. A simbologia V1,. . . ,V6, representa de forma simplificada as seis variáveis selecionadas para prosseguimento das análises, de modo que: V1 = Proporção do limite utilizado , V2 = Idade, V3 = Quantidade de atrasos de 30 a 59 dias, V4 = Proporção da renda comprometida com dívida, V5 = Quantidade de empréstimos ativos e V6 = Quantidade de empréstimos imobiliários. 80

4.6 Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para a variável Proporção do limite utilizado. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Idade. 81

4.7 Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para a variável Quantidade de atrasos entre 30 e 59 dias. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Proporção da dívida em relação à renda. 82

4.8 Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para a variável Quantidade de empréstimos ativos. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Quantidade empréstimos imobiliários. 82

Lista de Tabelas

3.1 Matriz de confusão para classificação binária no contexto de análise de crédito: n é o número de clientes no conjunto de teste, i_I é o número de clientes inadimplentes classificados corretamente, i_A é o número de clientes inadimplentes classificados incorretamente, a_I é o número de clientes adimplentes classificados incorretamente, a_A é o número de clientes adimplentes classificados corretamente, i é o número de clientes inadimplentes no conjunto de teste, a é o número de clientes adimplentes no conjunto de teste, I é o número de clientes classificados como inadimplentes e A é o número de clientes classificados como adimplentes. 64

4.1 Resultados da aplicação dos classificadores no conjunto de teste do conjunto de dados Inadimplência de clientes de cartão de crédito. Nas colunas, estão representados as medidas utilizadas para medir o desempenho dos classificadores, em que ACC = Acurácia, SEN = Sensibilidade, VPP = Valor Preditivo Positivo, VPN = Valor Preditivo Negativo, GMedia = G-Média e MCC = Coeficiente de Correlação de Matthews. Nas linhas estão dispostos os classificadores, os quais foram simbolizados pelas siglas, SVM = Máquinas de Vetores Suporte, RL = Regressão Logística, FA = Floresta Aleatória e SVMSC = Máquinas de Vetores Suporte Sensível ao Custo. Os resultados estão apresentados em dois blocos, dados originais, em que o ajuste foi realizado no conjunto de treinamento sem nenhuma alteração e SMOTE, no qual o ajuste foi realizado no conjunto de treinamento balanceado por tal técnica. A marcação em negrito, destaca qual dos três classificadores (SVM, RL e FA) teve melhor desempenho para as medidas avaliadas em cada um dos blocos. 76

4.2 Resultados da aplicação dos classificadores, abordados no estudo, no conjunto de teste do conjunto de dados Inadimplência de crédito. Nas colunas estão representados as medidas utilizadas para medir o desempenho dos classificadores, em que ACC = Acurácia, SEN = Sensibilidade, VPP = Valor Preditivo Positivo, VPN = Valor Preditivo Negativo, GMedia = G-Média e MCC = Coeficiente de Correlação de Matthews. Enquanto que nas linhas estão representados os classificadores, no qual foram simbolizados pelas respectivas siglas, SVM = Máquinas de Vetores Suporte, RL = Regressão Logística, FA = Floresta Aleatória e SVMSC = máquinas de vetores suporte sensível ao custo. Os resultados estão apresentados em dois blocos, dados originais, em que o ajuste foi realizado no conjunto de treinamento sem nenhuma alteração e SMOTE, no qual o ajuste foi realizado no conjunto de treinamento balanceado por tal técnica. A marcação em negrito, destaca qual dos três classificadores (SVM,RL e FA) tive melhor desempenho para as medidas avaliadas em cada um dos blocos. 84

Sumário

1 Introdução **25**

2 Análise de crédito **29**

2.1 Crédito 29

2.1.1 Operações de crédito 30

2.1.2 Política de crédito 31

2.1.3 Concessão de crédito 32

2.1.4 Limite de crédito 32

2.2 Análise de crédito 33

2.2.1 Análise subjetiva ou financeira 34

2.2.2 Análise objetiva ou estatística 35

2.2.3 Risco de crédito 36

2.2.4 Classificação de crédito 37

3 Classificação de crédito **39**

3.1 Classificação 39

3.1.1 Hiperplano 40

3.1.2 Classificação usando um hiperplano de separação 42

3.1.3 Classificador de margem máxima 45

3.1.4 Construção do classificador de margem máxima 47

3.1.5 Classificador de vetores suporte 50

3.1.6 Máquinas de vetores suporte 53

3.1.7 Processo de otimização 57

3.1.8 Máquinas de vetores suporte e dados desbalanceados 58

3.1.9 Máquinas de vetores suporte sensível ao custo 60

3.1.10 SMOTE 61

3.2 Medidas de performance 63
3.2.1 Matriz de confusão 64
3.2.2 Medidas de performance baseadas na matriz de confusão 65
3.2.3 Medidas de performance baseadas em análises individuais 66

4 Aplicações em dados reais **69**
4.1 Inadimplência de clientes de cartão de crédito 69
4.1.1 Análise descritiva e exploratória dos dados 71
4.1.2 Resultados 75
4.2 Inadimplência de crédito 78
4.2.1 Análise descritiva e exploratória dos dados 80
4.3 Resultados 83
4.4 Discussão 85

5 Considerações Finais **89**
5.1 Estudos Futuros 89
5.2 Considerações Finais 90

Referências Bibliográficas **92**

Capítulo 1

Introdução

Uma pontuação de crédito é um número calculado com base no histórico financeiro de quem solicita o serviço, utilizado para mensurar os riscos de conceder crédito ao cliente (Investopedia, 2021). Nesse sentido, a pontuação de crédito, também conhecida como classificação de crédito, é uma análise que as instituições financeiras realizam para prever a probabilidade de inadimplência dos solicitantes de produtos financeiros com base em determinadas características tais como idade, profissão, estado civil, endereço, renda, fluxos financeiros, registros de pagamento etc (Thomas *et al.*, 2002). Dessa forma, a análise de crédito é fundamental para as tomadas de decisões de concessão de crédito das instituições financeiras, uma vez que através dela é possível identificar os clientes que pertencem a um perfil que apresenta maior probabilidade de se tornar um devedor.

A necessidade da criação de modelos estatísticos para a classificação de crédito surge em 1960 quando os cartões de crédito começaram a ser introduzidos. Um bom modelo estatístico de classificação de crédito é capaz de agrupar efetivamente os clientes em grupos adimplentes e inadimplentes. Assim, quanto mais eficiente for o modelo, mais custos podem ser economizados pela instituição financeira. O trabalho pioneiro para discriminar entre solicitações de crédito boas (adimplentes) e ruins (inadimplentes) usando o método estatístico deve-se a Durand (1941). Desde então, diversos métodos de aprendizado de máquina e outros métodos estatísticos foram propostos para lidar com a questão da classificação dos solicitantes de créditos em bons e maus pagadores. Dentre eles, os mais populares são a análise de discriminante linear (Altman, 1968), regressão logística (Wiginton, 1980), redes neurais (Altman *et al.*, 1994; Desai *et al.*, 1996), árvores de decisão (Arminger *et al.*, 1997), e máquinas de vetores suporte (Baesens *et al.*, 2003). Em particular, máquinas de vetores suporte são frequentemente escolhidas como o classificador

base em algoritmos de classificação de crédito ver, por exemplo, Zhou *et al.* (2010); Li *et al.* (2013). Nesse contexto, Baesens *et al.* (2003) foi o primeiro a construir modelos de classificação de crédito com máquinas de vetores suporte e comparar a sua performance com outros métodos da literatura. Nesse trabalho, eles concluíram que as máquinas de vetores suporte possuem uma boa performance quando comparadas com os outros métodos. Recentemente, alguns estudos de revisão (Alaka *et al.*, 2018; Moro *et al.*, 2016) tem identificado as máquinas de vetores suporte como o classificador base escolhido por muitos pesquisadores para o desenvolvimento de modelos de crédito.

Máquinas de vetores suporte (SVM, do inglês *Support Vector Machine*) foi introduzida por Vapnik (1998), no contexto da teoria de aprendizagem estatística. É um conceito que toma como entrada um conjunto de dados e prediz, para cada nova entrada dada, qual de duas possíveis classes essa entrada pertence. Portanto, SVM é um classificador binário não-probabilístico. Essencialmente, essa técnica constrói um hiperplano no espaço das variáveis que estão sendo consideradas no estudo através de algum mapa, escolhido *a priori*. Esse hiperplano divide o espaço das variáveis em dois subconjuntos, de tal forma que a separação entre os subconjuntos seja tão ampla quanto possível. As novas entradas são, então, mapeadas no espaço das variáveis e preditas como pertencente a uma das duas possíveis classes baseada em qual subconjunto são colocadas.

Uma situação comum em aplicações de classificação de crédito é que os dados coletados geralmente apresentam um desbalanceamento das classes, ou seja, o número de clientes inadimplentes (classe minoritária) é muito menor do que o número de clientes adimplentes (classe majoritária). Um modelo ideal para pontuação de crédito deve ter um bom desempenho tanto para clientes pertencentes a classe majoritária (clientes adimplentes) quanto para aqueles da classe minoritária (clientes inadimplentes). Os algoritmos usuais de aprendizagem são, em geral, a favor da classe majoritária e, geralmente, apresentam baixo desempenho na classificação de indivíduos da classe minoritária (Wang *et al.*, 2015). Nesse sentido, uma alternativa é considerar conjuntos de treinamento equilibrados, obtidos a partir de algum pré-processamento de dados, ou utilizar um classificador que consiga captar e mitigar o impacto do desbalanceamento de classes.

Neste trabalho, propomos estudar a performance das máquinas de vetores suporte na classificação de crédito utilizando como conjunto de treinamento tanto dados balanceados quanto dados desbalanceados. Para tornar o conjunto de dados balanceados, iremos seguir duas vertentes apontadas na literatura, que são os métodos externos de balanceamento

e os métodos internos. Os métodos externos consistem em realizar uma modificação no conjunto de treinamento de modo a tornar as classes equilibradas, para esse fim, aplica-se um pré-processamento no conjunto de treinamento. Os métodos de pré-processamento frequentemente utilizados são a sobreamostragem e a subamostragem das classes. Na sobreamostragem realiza-se a duplicata dos dados que estão menos presente no conjunto de treinamento, isto é, na classe minoritária (clientes inadimplentes), enquanto que na subamostragem o processo inverso é realizado, descartando informações da classe mais populosa (clientes adimplentes). Em seu estudo Chawla *et al.* (2002) argumenta que tais métodos podem trazer ruídos para análise, dado que na sobreamostragem estamos duplicando as informações já existente e na subamostragem estamos perdendo informações, com a retirada de observações. Nesse sentido, os autores propuseram um método de sobreamostragem que cria novas observações sintéticas a partir de uma observação real com o incremento de uma perturbação. Desse modo, a classe minoritária passa a ser populada por novas observações que são distantes daquelas já existentes, porém muito semelhantes. Essa técnica foi denominada como SMOTE (*Synthetic Minority Oversampling Technique*) e a utilizaremos em nosso estudo para balanceamento dos conjuntos de treinamento.

Uma outra abordagem para superar o viés ocasionado pelo desequilíbrio de classes é a partir dos métodos internos, que consistem em aplicar modificações no algoritmo dos classificadores, de modo a tornar estes mais calibrados para lidar com dados desbalanceados. Nessa monografia, abordaremos o classificador de máquinas de vetores suporte sensível ao custo, uma técnica proposta por Veropoulos *et al.* (1999) que permite atribuir custos distintos às classes em análise, de modo que errar a classificação de uma observação da classe minoritária seja mais penalizado do que errar uma classificação da classe majoritária. Tal técnica, é considerada um método interno, devido ao fato de alterar a proposta inicial do algoritmo de máquina de vetores suporte, que considera o custo igual para ambas as classes.

O estudo comparativo da performance das máquinas de vetores suporte no cenário com conjunto de treinamento balanceado com aquele com conjunto de treinamento desbalanceado será realizado em termos de certas medidas de performance. Esse estudo comparativo será realizado em um cenário de dados reais, no qual abordaremos dois conjuntos de dados, um com desequilíbrio de classe moderado e o outro mais severo. Além disso, para averiguar as conclusões obtidas por Baesens *et al.* (2003) iremos comparar o desempenho do classificador de máquinas de vetores suporte com outros classificadores constantemente

citados na literatura, como o classificador de floresta aleatória e regressão logística. Tais comparações serão realizadas no cenário de conjuntos de treinamento desbalanceados e balanceados via SMOTE.

Este trabalho está organizado da seguinte maneira. No próximo capítulo, apresentamos o funcionamento do processo de análise de crédito realizado pelas instituições financeiras. No Capítulo 3, estudamos a metodologia de máquinas de vetores suporte que será utilizada para classificação de crédito, explicaremos como será realizado o balanceamento do conjunto de treinamento e abordaremos as definições teóricas do classificador de máquinas de vetores suporte sensível ao custo. As medidas que utilizaremos para avaliar a performance do classificador de crédito também são apresentadas no Capítulo 3. No Capítulo 4 apresentamos os resultados obtidos, assim como uma discussão mais detalhada acerca de tais resultados. O Capítulo 5 encerra esta monografia com algumas considerações finais, conclusões e sugestões para estudos futuros.

Capítulo 2

Análise de crédito

2.1 Crédito

Crédito é um termo que tem origem na expressão "crer", e pode ser traduzido como uma relação de confiança. O termo crédito pode ser definido de diversas maneiras dependendo do contexto em que é observado, entretanto, sob o aspecto financeiro é possível defini-lo como:

> "Conceder a um tomador recursos financeiros para fazer frente a despesas ou investimentos, ou para qualquer outra finalidade, atrelados a um retorno esperado e pré-determinado pelo credor (BLATT, 1999)."

Dessa forma, crédito consiste na entrega de um valor presente por uma entidade financeira a um terceiro por um período determinado confiando que este, futuramente, seja capaz de retornar o que lhe foi confiado. Aquele que recebe o crédito (tomador do crédito) deve devolver o valor concedido acrescido de uma remuneração, previamente estabelecida, como forma de compensação pelo valor obtido.

Em um sistema econômico as operações de crédito tem um papel fundamental no desenvolvimento e crescimento desse sistema. Segundo BRIGHAM *et al.* (2005), a comercialização de crédito por instituições financeiras e empresas é um relevante fator que impulsiona e impacta todos os setores da economia. Dentre os impactos econômicos e sociais gerados pela concessão de crédito podemos citar a possibilidade das empresas inovarem e executarem projetos para os quais não possuíam recursos suficientes, podendo resultar no aumento de sua atividade produtiva, o aumento do consumo que impacta a demanda dos bens consumidos, além de disponibilizar recursos financeiros para aquisição

de moradia ou bens pela sociedade como um todo. A finalidade de crédito está diretamente relacionada às necessidades do cliente tomador, portanto, é necessário que a linha de crédito oferecida pela instituição financeira seja compatível com a situação financeira, patrimonial, necessidade de financiamento e sua capacidade de amortização do valor concedido. Entretanto, apesar da concessão de crédito ser propulsor de desenvolvimento social e econômico, a mesma é um fator que contribui para o endividamento de empresas ou pessoas físicas, e, muitas vezes, prejudicial durante um processo inflacionário.

2.1.1 Operações de crédito

Segundo o Banco Central do Brasil (BACEN) as operações de crédito podem ser classificadas quanto as modalidades de crédito que uma instituição financeira comercializa, como também de acordo com a autonomia dos recursos que são transacionados. A depender do tipo de recurso que é ofertado empregam-se diferentes modalidades de crédito.

A autonomia dos recursos podem ser classificadas de duas maneiras: operações de crédito com recursos livres e operações de crédito com recursos direcionados. Segundo a definição apresentada em BACEN (2019), operações de crédito com recursos direcionados correspondem a operações regulamentadas pelo CMN (Conselho Monetário Nacional) ou vinculadas a recursos orçamentários destinados à produção e ao investimento de médio e longo prazos aos setores imobiliário, rural e de infraestrutura. Geralmente o capital é oriundo de fundos e programas públicos facilitados pelo governo ou de recursos capitados pela própria instituição financeira, como a caderneta de poupança. As principais modalidades ofertadas com esse tipo de recurso são: capital de giro com recursos do BNDES, crédito imobiliário, rural ou de infraestrutura vinculados a taxas regulamentadas ou de mercado.

Já os recursos livres são definidos como contratos de financiamentos e empréstimos com taxas de juros a ser definidas pela instituição financeira e o solicitante de crédito. Nesse tipo de operação a instituição tem autonomia para destinação dos recursos captados, sendo as modalidades mais comercializadas: empréstimos para aquisição de veículos, empréstimos para aquisição de outros bens, empréstimos para capital de giro, compras à vista, parceladas ou rotativas do cartão crédito, cheque especial e crédito pessoal.

2.1.2 Política de crédito

Políticas são criadas por instituições financeiras como instrumentos que compõem padrões e regras de decisões para solucionar problemas semelhantes e tem o objetivo de orientar a decisão de crédito, levando em consideração os objetivos desejados e estabelecidas pela instituição concedente do crédito. Uma política de crédito eficaz deve assegurar que a instituição financeira alcance os resultados financeiros pretendidos com alta confiabilidade, ou seja, deve ajustar risco e minimizar perdas, paralelamente ao alcance da meta estabelecida. Para atingir esse objetivo as políticas de crédito devem ser estratégicas e incluírem:

a) Critérios de desempenho: que definem parâmetros que avaliam o cumprimento das políticas e dos comportamentos por ela estabelecidos;

b) Procedimentos: especificam as atividades que devem ser desempenhadas para garantir que os padrões definidos sejam consistentes diante de diferentes situações e adversidades.

Portanto, a política de crédito de uma empresa ou instituição deve ser uma regulamentação base a ser seguida e aplicada em situações em que se faz necessário, sendo esta guiada pelas crenças e objetivos estratégicos (TSURU e CENTA, 2009). Uma política de crédito deve seguir normas estabelecidas por autoridades monetárias. Entretanto, tanto as políticas quanto os procedimentos devem ser ajustados de acordo com as perspectivas táticas de cada instituição financeira, devendo estar em concordância com sua concepção de crédito.

A partir do estabelecimento da Política de Crédito de uma instituição, são então definidos alguns parâmetros e indicadores, como as taxas de juros, os prazos de recebimentos, limite do crédito, as garantias necessárias e o risco que a mesma está disposta a assumir nas operações de crédito. Na estruturação de uma política de crédito muitos fatores devem ser considerados, uma vez que a mesma não deve ser inteiramente restritiva de forma a dificultar a concessão do crédito, por outro lado não deve ser totalmente liberal se expondo ao alto risco de inadimplência. Dessa forma é aconselhável que uma política de crédito seja definida de forma equilibrada preservando os recursos da instituição e viabilizando a construção de uma carteira de clientes sólida e lucrativa. Dessa maneira, pontuações ou classificações de crédito via métodos de aprendizado estatístico têm sido uma das grandes ferramentas dentro da definição das políticas de crédito.

2.1.3 Concessão de crédito

Dentro do processo operacional de análise de crédito, a concessão do crédito consiste na etapa em que a instituição financeira irá definir se concederá ou não crédito a parte solicitante. São três os elementos fundamentais que devem ser considerados durante a operação de concessão de crédito: segurança, liquidez e rentabilidade (BLATT, 1999). A segurança consiste no risco inerente ao processo de concessão de crédito, o qual é reduzido após análise prévia e minuciosa da ficha cadastral e demonstrativos contábeis do cliente. A liquidez se refere a capacidade de pagamento do crédito concedido acrescido dos encargos contratuais acordados, caso o cliente tomador de crédito dependa da aprovação de crédito por outro credor para liquidar a operação, a condição de liquidez não é satisfeita. Por fim, a rentabilidade descreve a obtenção de lucro por meio da operação de crédito, além de ser segura e líquida, a operação deve ser rentável a fim de não comprometer a liquidez do credor.

Em geral, as instituições financeiras são criteriosas na tomada de decisão de crédito, uma vez que o não retorno de uma operação de crédito representa perda do montante emprestado e prejuízo para o credor. Dessa forma, a concessão de crédito deve estar fundamentada em criteriosas análises, seguindo o que está definido na política de crédito de cada instituição. Segundo Neto e Silva (1997), a concessão de crédito deve ser uma resposta individual da instituição financeira para o cliente tomador do crédito, caso o cliente satisfaça as condições pré-estabelecidas pela concedente do crédito, o crédito pode ser então concedido.

2.1.4 Limite de crédito

A decisão de concessão de crédito pela instituição financeira pode ser restrita a necessidade de um cliente requerendo uma análise mais específica, ou pode ser mais abrangente, como é o caso do limite de crédito. O limite de crédito constitui uma ação realizada pelas instituições financeiras na qual é definido um limite para atendimento do cliente ou conjunto de empresas de uma área de negócios. O estabelecimento de um limite de crédito tem a finalidade de conferir maior velocidade às decisões de crédito, e assim, proporcionar a instituição financeira maior competitividade, uma vez que não se faz necessária a análise caso a caso, desde que os clientes se enquadrem nas condições definidas. Além de conferir agilidade, o limite de crédito pode promover maior uniformidade no atendimento

das necessidades dos clientes ou das empresas atendidas pela intuição financeira.

O estabelecimento de limite de crédito necessita de uma revisão periódica do mesmo, sendo aconselhável que o prazo dos limites estabelecidos não seja superior a um ano. A periodicidade da revisão do limite de crédito será dependente das condições gerais dos negócios, da classificação de risco do cliente/ empresa e do contexto macroeconômico. Além disso, apesar dessa modalidade de concessão de crédito não requerer uma análise específica do tomador de crédito, a instituição financeira necessita conhecer as necessidades de seus clientes, identificar quais dos seus produtos que mais atendem tais necessidades e, então, definir o limite de crédito a ser disponibilizado. A análise prévia dos clientes tomadores de crédito, evita que a instituição concedente do crédito não sub ou superestime os limites de créditos oferecido, orientando para propostas mais estruturadas, competitivas, precisas e de menor risco.

Portanto, a fixação do limite de crédito deve ser sustentada por três pilares: I) as necessidades do cliente, (II) o risco de crédito apresentado pelo cliente, e (III) a Política de Crédito da instituição. Nesse processo, a instituição financeira deve manter o equilíbrio entre o quanto o cliente solicita de crédito com o quanto a instituição pode e/ou deve oferecer de crédito. Nesse sentido, uma política restritiva tende a dificultar a liberação do crédito, prejudicando na comercialização dos serviços da instituição, por outro lado, uma política muito abrangente a expõe ao risco de inadimplência. Para tal, diversos fatores devem ser analisados, como o tipo de modalidade de crédito oferecida, o prazo estipulado e a expectativa de retorno do montante fornecido, por exemplo.

2.2 Análise de crédito

A análise de crédito compreende na avaliação da capacidade do solicitante de crédito arcar com seus compromissos futuros, retornando ao credor a quantia que lhe foi cedida. Essa avaliação compreende da aplicação de técnicas estatísticas, financeiras e análises subjetivas. Logo, essa é uma etapa crucial na realização de uma operação de crédito, pois é a partir da análise de crédito que a instituição financeira irá se fundamentar na decisão da concessão de seus recursos aos tomadores de créditos. Em concordância com a aprovação do crédito, tal análise também auxilia na definição dos riscos envolvidos na operação e a que condições o crédito deve ser liberado (prazo, garantias, taxa de juros, etc).

2.2.1 Análise subjetiva ou financeira

O processo de análise subjetiva envolve decisões individuais quanto à concessão ou recusa do crédito por parte da instituição financeira. Nesse processo, a tomada de decisão é baseada na experiência adquirida, na disponibilidade de informações e na sensibilidade de cada analista quanto à viabilidade da liberação de crédito. Entretanto, tais decisões não devem ser tomadas pelo analista de forma aleatória, baseado apenas em uma decisão momentânea. Para auxiliar na tomada de decisão dos analistas, existem documentações e critérios pré-definidos para serem seguidos. As informações necessárias para uma análise subjetiva da capacidade financeira dos tomadores de crédito são tradicionalmente conhecidas como os 6 Cs: Caráter, Capacidade, Capital, Colateral, Condições gerais e Conglomerado. De modo geral podemos descrevê-los como:

- **Caráter:** representa a integridade do cliente no mercado financeiro. Ou seja, baseado no histórico de pagamento e existências de restrições no mercado é possível investigar sua idoneidade em pagar suas dívidas. Através da consulta em órgãos de proteção de crédito, é possível avaliar as operações financeiras realizadas pelo tomador de crédito, seja ele pessoa física ou jurídica;

- **Capacidade:** se refere a habilidade do cliente em converter investimentos em receitas, traduzido pelo potencial do cliente em saldar os créditos recebidos. No caso de pessoas físicas, a capacidade se refere a habilidade do indivíduo de gerir sua vida pessoal financeira, sua estabilidade financeira, empreendimentos e respectivos sucessos, além de informações pessoais como idade, estado civil, cônjuge entre outras. Considerando pessoas jurídicas são analisados experiência dos administradores, capacidade produtiva da empresa, instalação física, grau tecnológico, competitividade de mercado, entre outros. A análise de capacidade é bastante subjetiva, porém almeja analisar o fato de se a pessoa ou empresa sabe gerir bem suas finanças e economias;

- **Capital:** faz menção às condições econômico-financeiras do cliente, analisando suas condições de negócios, ramo de atividade, volume de bens e direitos disponíveis para cumprimento das obrigações financeiras contratuais. É realizada através de demonstrativos financeiros e contábeis da pessoa física ou da empresa solicitante do crédito;

- **Colateral:** trata-se de garantias oferecidas pelo devedor que o confiram maior segurança de crédito. Em algumas situações é utilizado para contrabalancear os fatores que desfavorecem o cliente analisado quanto ao capital e a capacidade, e assim atenuar eventuais riscos;

- **Condições Gerais:** se refere ao contexto macroeconômico, ou seja, fatores externos ao tomador de créditos e que fogem do seu controle. Os principais aspectos que afetam a análise dessa variável são: informações referentes ao ramo mercado e produto de atuação, conjuntura política do país, qualidade de produtos e serviços, competitividade, necessidade de subsídios do governo, entre outros.

- **Conglomerado:** avalia a situação financeira dos participantes do mesmo grupo econômico. Nesse contexto, não basta conhecer a situação econômica de uma empresa isolada, mas é necessário realizar a análise financeira de sua controladora, controladas, interligadas e coligadas. Em relação a pessoas físicas, o conglomerado inclui a análise de crédito de cônjuges, dependentes, garantidores de crédito, referências ou outras.

2.2.2 Análise objetiva ou estatística

A concessão de crédito é uma decisão tomada em um cenário de incertezas, no qual confia-se que o tomador de crédito será capaz de cumprir com as condições determinadas no ato da concessão. Nesse sentido, é do interesse do credor estimar os riscos e as chances do cliente ruir em perdas, para assim, ter mais convicções na decisão de crédito. A análise objetiva é uma poderosa ferramente para esse fim, baseada em metodologias estatísticas tem como objetivo contabilizar resultados matemáticos que auxiliam na tomada de decisão das instituições concedentes de crédito.

De modo geral, para realizar as análises são utilizados modelos estatísticos desenvolvidos com base em dados históricos, que são compostos por informações cadastrais, financeiras, patrimoniais e de idoneidade dos clientes. E a partir das análises geram-se métricas, medidas de probabilidade ou grupos de classificações que auxiliarão na decisão de concessão de crédito de uma instituição financeira.

Dentre as técnicas objetivas de gestão do risco de crédito, destacamos o credit scoring, behavioural scoring, ratings e previsão de insolvência. Os modelos de credit scoring são utilizados para classificar clientes sujeitos à obtenção de crédito em grupos de risco, essa

classificação ocorre por meio de pontuações que são calculadas para cada cliente ou grupos de clientes. A metodologia de behavioural scoring é considerada uma subclasse de credit scoring, em que geralmente é aplicada para avaliação comportamental de crédito, no sentido de gerenciar limites de crédito, ações preventivas, entre outras estratégias, para clientes que já tem algum relacionamento com a instituição.

A metodologia de Rating é uma classificação de risco de crédito que pode ser atribuída a um país, uma empresa, uma pessoa, ou a uma operação de crédito, no qual é apresentado por meio de um código que fornece uma graduação de risco. Geralmente, essa classificação é representada por escalas de letras ou números, por exemplo numa escala de variação de AA a H, em que empresas classificadas com AA apresentam menores riscos de crédito. Por fim, os modelos de previsão a solvência ou insolvência, são modelos desenvolvidos com intuito de quantificar a chance de empresas se tornarem insolventes. Empresas insolventes são empresas que não tem condições de cumprir com suas obrigações financeiras.

2.2.3 Risco de crédito

Risco pode ser definido como probabilidade de que algo almejado ou esperado não ocorra, por conta de um acontecimento indesejável, incerto, ou que a ocorrência não depende exclusivamente do desejo da parte interessada. Analogamente, definimos risco de crédito como a possibilidade do tomador de crédito não cumprir com as condições estabelecidas na concessão de crédito, em outras palavras é a chance de inadimplência em uma operação de crédito.

Ao realizar a concessão de crédito, a intermediária financeira assume o risco e os benefícios que a transação envolve. Acontecimentos adversos e imprevistos, como por exemplo os resultantes de recessão econômica, podem impactar nas fontes de renda de pessoas físicas e empresas, reduzindo a probabilidade de pagamento do valor concedido. Segundo WESLEY (1993), dois fatores são determinantes para o risco de inadimplência por clientes tomares: (i) a fraca qualidade do processo de análise de crédito, e (ii) o agravamento da situação macroeconômica. Em contraste à situação macroeconômica que consiste em um fator externo, a análise de crédito para determinação dos riscos de inadimplência, representa um fator interno e que pode ser controlado pela instituição financeira. Portanto, a mesma deve ter alto nível de previsibilidade e confiança a fim de evitar perdas financeiras que impactam a liquidez e a captação de recursos no mercado financeiro pelas instituições financeiras.

2.2.4 Classificação de crédito

Classificação de crédito ou também conhecido pelo termo em inglês *credit scoring*, consiste na utilização de métodos matemáticos e estatísticos para classificação de candidatos a concessão de crédito. É partir dessa classificação que a instituição financeira irá decidir pela concessão ou não do crédito. Para construção dos modelos estatísticos é utilizado um conjunto de informações coletadas a cerca de clientes pertencentes a base cadastral da instituição, para que possa se formar um critério de análise quantitativo ou qualitativo das experiências observadas do passado, que irá servir como base para classificação de um novo cliente.

Dessa forma, o objetivo do processo de classificação de crédito é identificar o perfil e o comportamento financeiro dos indivíduos solicitantes de crédito, com base em suas informações básicas e histórico financeiro e, a partir desses levantamentos, tomar a decisão em relação à concessão do crédito. Algumas técnicas estatísticas já estão consolidadas no mercado de crédito, como regressão logística, análise de discriminante e árvores de decisões. Entretanto, notamos uma crescente busca por novas técnicas que resultem em melhores resultados e consequentemente melhore o poder de decisão para concessão de crédito de uma instituição financeira.

Umas dessas técnicas é a máquinas de vetores suporte, uma metodologia desenvolvida por Vapnik (1998) que vem ganhando cada vez mais espaço em estudos de crédito. Segundo estudos presentes na literatura, essa metodologia tem apresentado excelente performance (tem o mesmo sentido de desempenho) de classificação frente as demais técnicas presentes no mercado. Nesse sentido, a metodologia de máquinas de vetores suporte será abordada como tema de estudo dessa monografia, ao qual será apresentado, nos próximos capítulos, uma revisão da teoria e aplicação dos conceitos abordados.

Capítulo 3

Classificação de crédito

Conforme discutido no capítulo anterior, a concessão de crédito é um fator de suma importância para o crescimento da economia, uma vez que impulsiona o consumo e o investimento de capital em produtos financeiros que estão associados ao desenvolvimento de diversos setores da mesma. Todavia, a concessão de crédito representa uma atividade de risco podendo resultar em prejuízos para as instituições financeiras que disponibilizam seu capital sem executar uma análise delineada e segura de seu tomador de crédito. Anteriormente, neste trabalho, foram abordados procedimentos como criação de políticas de créditos e análises de riscos, sendo que a última envolve técnicas subjetivas e objetivas. No que se refere às técnicas objetivas, podemos destacar algumas metodologias estatísticas que podem aumentar a eficiência e a seguridade das decisões de concessão de crédito, as quais serão descritas em detalhes no decorrer deste capítulo.

3.1 Classificação

Máquinas de vetores suporte, do inglês *Support Vector Machines*, é um método de classificação proposto por Vapnik (1998). Tal técnica é uma generalização dos classificadores de margem máxima e vetores suporte, ao qual propõem uma regra de classificação baseada em um hiperplano de separação. Uma das principais suposições dos classificadores de margem máxima e vetores suporte é que o conjunto de dados seja linearmente separável ou que ao menos seja possível traçar uma fronteira linear entre as classes, o que restringe o uso desses classificadores em um cenário de dados reais. Nesse sentido, máquinas de vetores suporte aparece como uma solução para a classificação de dados cujas classes não são linearmente separáveis.

3.1.1 Hiperplano

Na geometria, dado p um número inteiro positivo, definimos um hiperplano em um espaço p-dimensional como sendo um subconjunto $p-1$-dimensional. Por exemplo, se um espaço é tridimensional, seus hiperplanos são os planos bidimensionais; enquanto se o espaço é bidimensional, seus hiperplanos são retas unidimensionais. Essa noção pode ser generalizada para quaisquer espaços vetoriais de quaisquer dimensões. Nesse sentido, um hiperplano é a generalização de plano para diferentes números de dimensões e para diferentes espaços vetoriais.

Definição 3.1 (Hiperplano) *Seja p um número natural não-nulo. Dado um vetor de parâmetros $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, o **hiperplano** $H_{\boldsymbol{\beta}}$ em $\mathbb{R}^p$ é definido como sendo o conjunto*

$$H_{\boldsymbol{\beta}} := \{\boldsymbol{x} \in \mathbb{R}^p : \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = 0\}.$$

Exemplo 3.2 *São exemplos de hiperplano:*

1. *Em $\mathbb{R}^3$, dado o vetor de parâmetros $\boldsymbol{\beta} \in \mathbb{R}^4$, um hiperplano é o conjunto*

$$H_{\boldsymbol{\beta}} := \left\{\boldsymbol{x} \in \mathbb{R}^3 : \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = 0\right\},$$

   ou seja, os pontos que satisfazem a equação de um plano.

2. *Em $\mathbb{R}^2$, dado o vetor de parâmetros $\boldsymbol{\beta} \in \mathbb{R}^3$, um hiperplano é o conjunto*

$$H_{\boldsymbol{\beta}} := \left\{\boldsymbol{x} \in \mathbb{R}^2 : \beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0\right\},$$

   ou seja, os pontos que satisfazem a equação de uma reta.

Dizer que, para um dado vetor de parâmetros $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, a equação

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = 0 \tag{3.3}$$

define um hiperplano $H_{\boldsymbol{\beta}}$, significa que qualquer vetor $\boldsymbol{x} \in \mathbb{R}^p$ que satisfaz a Equação (3.3), pertence ao hiperplano $H_{\boldsymbol{\beta}}$. Por outro lado, se o vetor $\boldsymbol{x}$ não satisfizer a Equação (3.3), então $\boldsymbol{x}$ não pertence ao hiperplano $H_{\boldsymbol{\beta}}$, e, consequentemente, $\boldsymbol{x}$ deve satisfazer

uma das desigualdades

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p > 0 \tag{3.4}$$

ou

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p < 0. \tag{3.5}$$

Logo, geometricamente, esse vetor $\boldsymbol{x}$ pertence a um dos lados originados pelo hiperplano $H_{\boldsymbol{\beta}}$. Portanto, o hiperplano $H_{\boldsymbol{\beta}}$ divide o espaço $\mathbb{R}^p$ em dois conjuntos disjuntos:

$$H^+_{\boldsymbol{\beta}} := \{\boldsymbol{x} \in \mathbb{R}^p : \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p > 0\}$$

e

$$H^-_{\boldsymbol{\beta}} := \{\boldsymbol{x} \in \mathbb{R}^p : \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p < 0\}.$$

Dessa forma, dado um vetor $\boldsymbol{x} \in \mathbb{R}^p$, ou ele pertence ao hiperplano $H_{\boldsymbol{\beta}}$ ou a um dos dois conjuntos disjuntos $H^+_{\boldsymbol{\beta}}$ e $H^-_{\boldsymbol{\beta}}$ definidos, respectivamente, pelas Equações (3.4) e (3.5).

Exemplo 3.6 *Seja $H_{\boldsymbol{\beta}}$ um hiperplano em $\mathbb{R}^2$ tal que*

$$\boldsymbol{\beta} = \begin{bmatrix} 0,2 \\ 0,85 \\ -1 \end{bmatrix}.$$

Assim,

$$\begin{aligned} H_{\boldsymbol{\beta}} &= \left\{\boldsymbol{x} \in \mathbb{R}^2 : 0,2 + 0,85x_1 - x_2 = 0\right\}, \\ H^+_{\boldsymbol{\beta}} &= \left\{\boldsymbol{x} \in \mathbb{R}^2 : 0,2 + 0,85x_1 - x_2 > 0\right\}, \\ H^-_{\boldsymbol{\beta}} &= \left\{\boldsymbol{x} \in \mathbb{R}^2 : 0,2 + 0,85x_1 - x_2 < 0\right\}, \end{aligned}$$

cuja representação no plano cartesiano é dada, respectivamente, pela reta em preto, os pontos em azul e os pontos em vermelho da Figura 3.1.

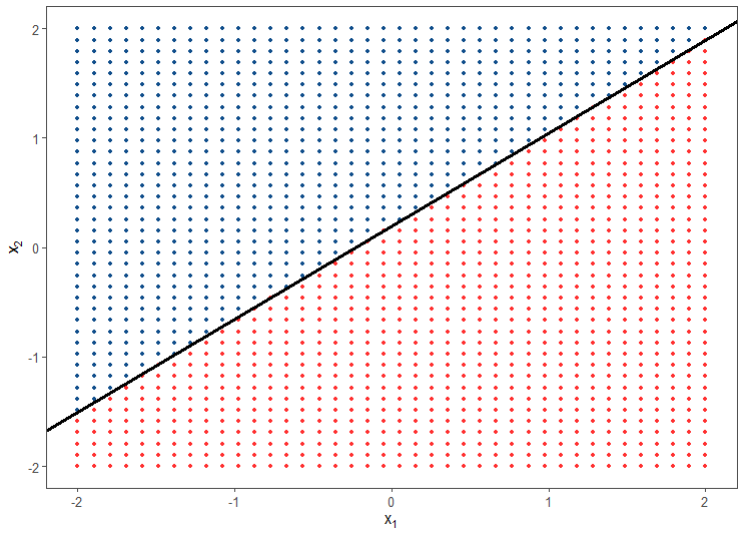

Figura 3.1: Hiperplano gerado pela equação $0,2+0,85x_1-1x_2=0$, representado pela reta em preto. Os pontos em azul representam o conjunto de pares ordenados que satisfazem $0,2+0,85x_1-1x_2<0$, e os pontos em vermelho representam o conjunto de pares ordenados que satisfazem $0,2+0,85x_1-1x_2>0$.

3.1.2 Classificação usando um hiperplano de separação

Dado um conjunto de dados $\mathcal{X}$, considere o subconjunto $Z=\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)\}$ de $\mathcal{X}$ como sendo um conjunto de treinamento, em que $\boldsymbol{x}_i\in\mathbb{R}^p$ é o vetor de características observadas associado ao indivíduo i e $y_i\in\{1,-1\}$ é a classificação do indivíduo i em adimplente ($y_i=-1$) ou inadimplente ($y_i=1$), em que i e n são números naturais tais que $1\leq i\leq n$. No contexto de classificação, o objetivo é encontrar uma regra, a partir do conjunto de treinamento, para qual seja possível classificar novas observações. Nesse sentido, queremos utilizar o conceito de hiperplano para obter tal regra de classificação.

Dessa forma, suponha que seja possível encontrar um hiperplano $H_{\boldsymbol{\beta}}$ que separe perfeitamente o conjunto de treinamento de acordo com a classe ao qual cada cliente pertence. Então, uma possível regra de classificação, baseada no hiperplano de separação $H_{\boldsymbol{\beta}}$, é obtido a partir da análise do sinal da função linear $f:\mathbb{R}^p\to\mathbb{R}$ tal que

$$f(\boldsymbol{x})=\beta_0+\beta_1x_1+\cdots+\beta_px_p, \tag{3.7}$$

em que $\boldsymbol{x}\in\mathbb{R}^p$ é o vetor de covariáveis observado e $\boldsymbol{\beta}\in\mathbb{R}^{p+1}$ é o vetor de parâmetros associado ao hiperplano $H_{\boldsymbol{\beta}}$ em questão. Nesse sentido, se $\boldsymbol{x}^*\in\mathbb{R}^p$ é o vetor de características observadas de um novo cliente, então

- quando $f(\boldsymbol{x}^*)>0$, temos $\boldsymbol{x}^*\in H_{\boldsymbol{\beta}}^+$ e, então, classificamos esse cliente como inadim-

plente, isto é, $y^* = 1$;

- quando $f(\boldsymbol{x}^*) < 0$, temos $\boldsymbol{x}^* \in H_{\boldsymbol{\beta}}^-$ e, então, classificamos esse cliente como adimplente, isto é, $y^* = -1$.

Exemplo 3.8 *Considere um conjunto de treinamento que possa ser separado perfeitamente através dos seguintes hiperplanos*

$$H_{\boldsymbol{\beta}_1} = \left\{\boldsymbol{x} \in \mathbb{R}^2 : 1 + 1x_1 - x_2 = 0\right\},$$
$$H_{\boldsymbol{\beta}_2} = \left\{\boldsymbol{x} \in \mathbb{R}^2 : 1 + 2x_1 - 0,5x_2 = 0\right\},$$
$$H_{\boldsymbol{\beta}_3} = \left\{\boldsymbol{x} \in \mathbb{R}^2 : 1 + 2x_1 - x_2 = 0\right\},$$

em que

$$\boldsymbol{\beta}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{\beta}_2 = \begin{bmatrix} 1 \\ 2 \\ -0,5 \end{bmatrix} \quad e \quad \boldsymbol{\beta}_3 = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}.$$

Na Figura 3.2, as retas em preto representam os hiperplanos definidos anteriormente, os pontos em azul são os pares ordenados com as características observadas dos clientes adimplentes e os pontos em vermelho são os pares ordenados com as características observadas dos clientes inadimplentes.

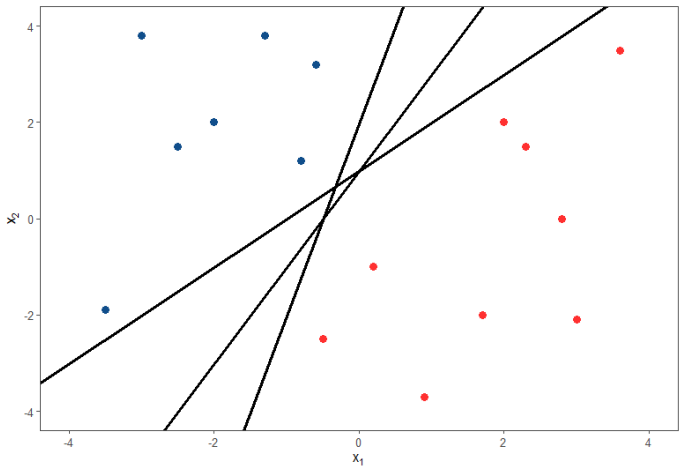

Figura 3.2: Hiperplanos $1 + x_1 - x_2 = 0$, $1 + 2x_1 - 0,5x_2 = 0$ e $1 + 2x_1 - x_2 = 0$ representados pelas retas em preto. Os pontos em azul representam o conjunto de pares ordenados associados aos clientes adimplentes e os pontos em vermelho representam o conjunto de pares ordenados associados aos clientes inadimplentes.

Exemplo 3.9 *Considere um conjunto de treinamento que possa ser separado perfeitamente através do hiperplano*

$$H_{\boldsymbol{\beta}} = \left\{ \boldsymbol{x} \in \mathbb{R}^2 : 0,8 + 1,2x_1 - x_2 = 0 \right\},$$

em que

$$\boldsymbol{\beta} = \begin{bmatrix} 0,8 \\ 1,2 \\ -1 \end{bmatrix}.$$

Nesse caso, a regra de classificação é dada pela função linear $f : \mathbb{R}^2 \to \mathbb{R}$ *tal que*

$$f(\boldsymbol{x}) = 0,8 + 1,2x_1 - x_2,$$

em que $\boldsymbol{x} \in \mathbb{R}^2$ *é o vetor de covariáveis observado.*

Na Figura 3.3 , a reta em preto representa o hiperplano definido anteriormente, os pontos em azul são os pares ordenados com as características observadas dos clientes adimplentes e os pontos em vermelho são os pares ordenados com as características observadas dos clientes inadimplentes.

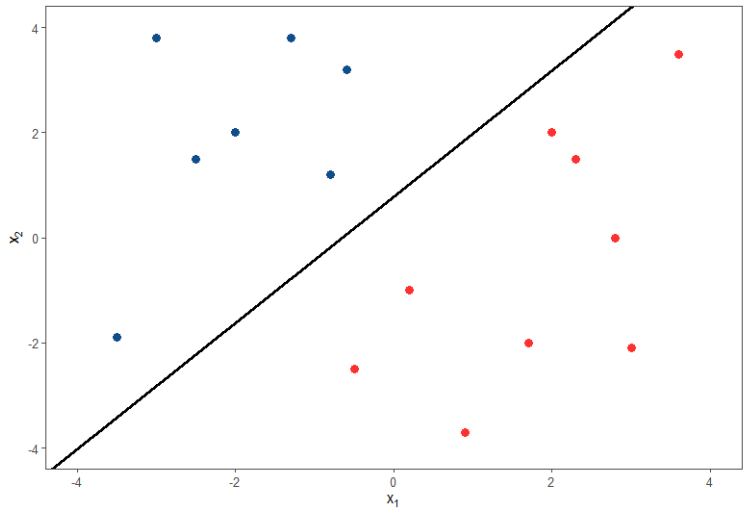

Figura 3.3: Hiperplano $0,8 + 1,2x_1 - x_2 = 0$ representado pela reta em preto. Os pontos em azul representam o conjunto de pares ordenados associados aos clientes adimplentes e os pontos em vermelho representam o conjunto de pares ordenados associados aos clientes inadimplentes.

Suponha que o vetor de características observadas

$$\boldsymbol{x}^* = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

de um novo cliente seja coletada. Como $f(\boldsymbol{x}^*) < 0$*, temos* $\boldsymbol{x}^* \in H_{\boldsymbol{\beta}}^-$ *e, então, classificamos esse cliente como adimplente, isto é,* $y^* = -1$*.*

3.1.3 Classificador de margem máxima

De modo geral, se for possível encontrar um hiperplano que separe perfeitamente o conjunto de treinamento em dois lados, ou seja, todas observações do conjunto de treinamento são divididas, de forma que nenhuma observação esteja classificada equivocadamente em um dos lados, então existirá diversos hiperplanos que podem ser obtidos que resultem nessa divisão. De fato, com pequenas alterações na inclinação ou na posição do hiperplano, iremos encontrar diferentes valores para o vetor de parâmetros $\boldsymbol{\beta}$, resultando em diferentes hiperplanos. Entretanto, é interessante que encontremos apenas um hiperplano, e que este seja aquele que resulte na melhor separação das classes, isto é, queremos encontrar o hiperplano tal que a distância das observações de fronteira das classes até o hiperplano seja máxima.. Esse hiperplano é conhecido como hiperplano de margem máxima.

Definição 3.10 (Margem) *Sejam* p *um número natural não-nulo e* $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ *um conjunto de treinamento, em que* $\boldsymbol{x}_i \in \mathbb{R}^p$ *é o vetor de características observadas associado ao indivíduo* i *e* $y_i \in \{1, -1\}$ *é a classificação do indivíduo* i *em adimplente* ($y_i = 1$) *ou inadimplente* ($y_i = -1$)*, em que* i *e* n *são números naturais tais que* $1 \leq i \leq n$*. Dado um hiperplano* $H_{\boldsymbol{\beta}}$*, definimos a* ***margem do hiperplano****, indicada por* $M(H_{\boldsymbol{\beta}})$*, como sendo a menor dentre todas as distâncias Euclidianas* d *de um vetor de características observado ao hiperplano, isto é,*

$$M(H_{\boldsymbol{\beta}}) = \min_{1 \leq i \leq n} d\left(\boldsymbol{x_i}, \operatorname*{proj}_{H_{\boldsymbol{\beta}}} \boldsymbol{x_i}\right),$$

em que $\operatorname*{proj}_{H_{\boldsymbol{\beta}}} \boldsymbol{x_i}$ *é a projeção ortogonal da observação* $\boldsymbol{x}_i$ *sobre o hiperplano* $H_{\boldsymbol{\beta}}$*.*

Nesse sentido, nosso objetivo é encontrar o hiperplano cuja margem seja máxima, ou seja, queremos o hiperplano que resulte na maior distância possível entre a margem e o

hiperplano. Dessa forma, podemos classificar uma nova observação com base no lado do hiperplano de margem máxima que essa se encontra, tal regra de classificação é conhecida como classificador de margem máxima.

Seja $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ um vetor de parâmetros que define o hiperplano ótimo, então o classificador de margem máxima que classifica novas observações $\boldsymbol{x}^* \in \mathbb{R}^p$ é baseado no sinal da função linear $f : \mathbb{R}^p \to \mathbb{R}$ tal que

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*.$$

As observações cujas distâncias até o hiperplano são as mínimas possíveis, dentre aquelas do conjunto de treinamento, são denominadas de **vetores suporte**, em virtude de pertencerem ao espaço p-dimensional e "suportarem" o hiperplano de margem máxima de modo que se forem alteradas suas coordenadas, o hiperplano ótimo também será alterado (Gareth *et al.*, 2013). Note que, nesse caso, os vetores suporte coincidem com as observações que pertencem à margem. Dessa forma, notamos que, para encontrarmos o hiperplano de margem máxima, precisamos apenas dos vetores suporte, sendo irrelevante as demais observações do conjunto de treinamento, pois a movimentação das mesmas não afetariam no hiperplano obtido.

Na Figura 3.4, temos um exemplo do hiperplano de margem máxima que separa perfeitamente as duas classes representadas pelos pontos em azul e vermelho.

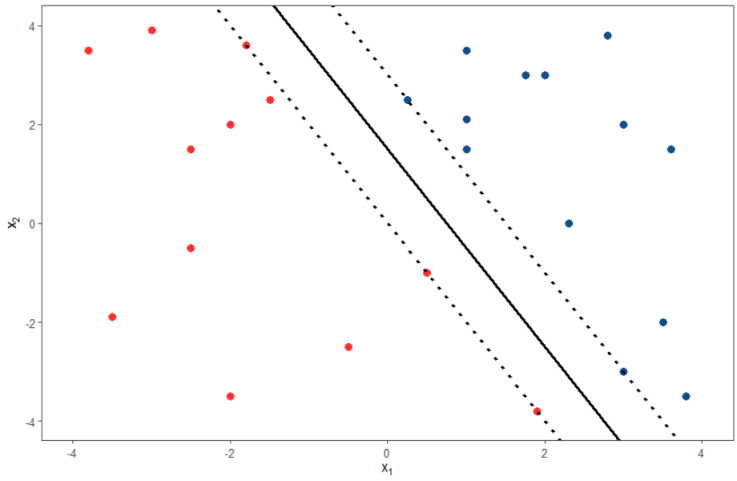

Figura 3.4: Os pontos em azul são os pares ordenados associados aos clientes adimplentes e os pontos em vermelho são os pares ordenados associados aos clientes inadimplentes. O hiperplano de margem máxima está representado pela linha em preto. Para este exemplo, a margem está representada pela linha pontilhada, que simboliza a menor distância de uma observação ao hiperplano. Os pontos azuis e vermelhos que estão sobre a linhas pontilhadas são os vetores suporte.

3.1.4 Construção do classificador de margem máxima

Para obtermos o classificador de margem máxima, dado um conjunto de treinamento $Z = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, em que $\boldsymbol{x}_i \in \mathbb{R}^p$ é o vetor de características observadas associado ao indivíduo i e $y_i \in \{1, -1\}$ é a classificação do indivíduo i em adimplente ou inadimplente, em que i e n são números naturais tais que $1 \leq i \leq n$, basta resolvermos o seguinte problema de otimização:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\text{maximizar}} \; M\left(H_{\boldsymbol{\beta}}\right) \tag{3.11}$$

$$\text{Sujeito a} \sum_{j=1}^{p} \beta_j^2 = 1, \tag{3.12}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M\left(H_{\boldsymbol{\beta}}\right) \quad \forall i = 1, \ldots, n, \tag{3.13}$$

em que $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ é o vetor de parâmetros e $M\left(H_{\boldsymbol{\beta}}\right)$ é a margem definida pelo hiperplano $H_{\boldsymbol{\beta}}$.

Note que a Equação (3.13) garante que as observações estarão classificadas corretamente, contanto que $M\left(H_{\boldsymbol{\beta}}\right)$ seja positivo. De fato, conforme abordado no classificador via hiperplano, para cada indivíduo $i \in \{1, 2, \ldots, n\}$, temos as seguintes situações:

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0 \quad se \quad y_i = 1$$

e

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0 \quad se \quad y_i = -1,$$

o que é equivalente a

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) > 0.$$

Dessa forma, se $M\left(H_{\boldsymbol{\beta}}\right)$ for maior do que zero, garantimos que para o hiperplano ótimo o classificador de margem máxima classifica todas as observações corretamente.

Além disso, podemos observar que a Equação (3.12) garante que a distância entre o i-ésimo vetor de características observado e a sua projeção ortogonal no hiperplano $H_{\boldsymbol{\beta}}$ é

dada por

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}). \tag{3.14}$$

Teorema 3.15 *Seja p um número natural não-nulo. Considere dados um hiperplano $H_{\boldsymbol{\beta}}$ em $\mathbb{R}^p$ e um conjunto de treinamento $Z = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, em que $\boldsymbol{x}_i \in \mathbb{R}^p$ é o vetor de características observadas associado ao indivíduo i e $y_i \in \{1, -1\}$ é a classificação do indivíduo i em adimplente ou inadimplente, em que i e n são números naturais tais que $1 \leq i \leq n$. Nessas condições, se*

$$\sum_{j=1}^{p} \beta_j^2 = 1,$$

então, para todo $i \in \{1, 2, \ldots, n\}$, é verdade que

$$d\left(\boldsymbol{x_i}, \operatorname*{proj}_{H_{\boldsymbol{\beta}}} \boldsymbol{x_i}\right) = y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}),$$

em que d é a distância Euclidiana e $\operatorname*{proj}_{H_{\boldsymbol{\beta}}} \boldsymbol{x_i}$ é a projeção ortogonal do vetor de características observadas $\boldsymbol{x_i}$ no hiperplano $H_{\boldsymbol{\beta}}$.

Demonstração. Suponha que $\boldsymbol{h} \in \mathbb{R}^p$ seja um ponto pertencente ao hiperplano $H_{\boldsymbol{\beta}}$, isto é,

$$\beta_0 + \beta_1 h_1 + \cdots + \beta_p h_p = 0.$$

Em outras palavras,

$$\beta_0 + \boldsymbol{\omega}^\top \boldsymbol{h} = 0, \tag{3.16}$$

em que

$$\boldsymbol{\omega} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

De fato, para qualquer $i \in \{1, 2, \ldots, n\}$,

$$\begin{aligned} d\left(\boldsymbol{x_i}, \underset{H_{\boldsymbol{\beta}}}{\text{proj}}\boldsymbol{x_i}\right) &= \|\text{proj}_{\boldsymbol{\omega}}(\boldsymbol{x}_i - \boldsymbol{h})\| \\ &= \left\| \frac{(\boldsymbol{x}_i - \boldsymbol{h})^\top \boldsymbol{\omega}}{\|\boldsymbol{\omega}\|^2} \boldsymbol{\omega} \right\| \\ &= \left| \boldsymbol{x}_i^\top \boldsymbol{\omega} - \boldsymbol{h}^\top \boldsymbol{\omega} \right| \frac{\|\boldsymbol{\omega}\|}{\|\boldsymbol{\omega}\|^2} \\ &= \frac{\left| \boldsymbol{x}_i^\top \boldsymbol{\omega} - \boldsymbol{\omega}^\top \boldsymbol{h} \right|}{\|\boldsymbol{\omega}\|}. \end{aligned}$$

.

Da Equação (3.16), segue que

$$-\boldsymbol{\omega}^\top \boldsymbol{h} = \beta_0.$$

Logo,

$$d\left(\boldsymbol{x_i}, \underset{H_{\boldsymbol{\beta}}}{\text{proj}}\boldsymbol{x_i}\right) = \frac{\left| \boldsymbol{x}_i^\top \boldsymbol{\omega} + \beta_0 \right|}{\|\boldsymbol{\omega}\|}.$$

Por hipótese,

$$\|\boldsymbol{\omega}\| := \sum_{j=1}^{p} \beta_j^2 = 1.$$

Portanto,

$$d\left(\boldsymbol{x_i}, \underset{H_{\boldsymbol{\beta}}}{\text{proj}}\boldsymbol{x_i}\right) = \left| \boldsymbol{x}_i^\top \boldsymbol{\omega} + \beta_0 \right| = y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}).$$

□

Dessa forma, as restrições (3.12) e (3.13) afirmam que a distância de qualquer vetor de características do conjunto de treinamento até o hiperplano ótimo deve ser maior ou igual à margem associada a esse hiperplano, o que garante que cada observação estará classificada do lado correto do hiperplano. Logo, o classificador de margem máxima é resultante da obtenção do vetor $\boldsymbol{\beta}$ que maximiza o valor da margem $M(H_\beta)$. Realizando algumas transformações em (3.11) - (3.13), podemos chegar em um problema de otimização convexa, que pode ser resolvido a partir do método de multiplicadores de Lagrange, que para maiores detalhes sugerimos que o(a) leitor(a) veja Ng (2000).

3.1.5 Classificador de vetores suporte

O classificador de margem máxima discutido na seção anterior tem como suposição, que o conjunto de dados seja linearmente separável, ou seja, definindo um hiperplano conseguimos separar perfeitamente os clientes em duas classes. Entretanto, de modo geral, trabalhamos com conjunto de dados que não conseguimos encontrar tal hiperplano, veja um exemplo na Figura 3.5. Nesse caso, não é possível utilizar o classificador de margem máxima, pois não conseguimos obter uma solução para o problema de otimização quando $M\left(H_{\boldsymbol{\beta}}\right) > 0$ (Gareth *et al.*, 2013). Uma possível alternativa para a classificação de dados quando não for possível obter o hiperplano ótimo é utilizar um classificador mais flexível, no qual é permitido que algumas observações extrapolem o limite da margem e até mesmo do hiperplano. O classificador cuja regra de classificação apresenta tal característica é denominado classificador de vetores suporte.

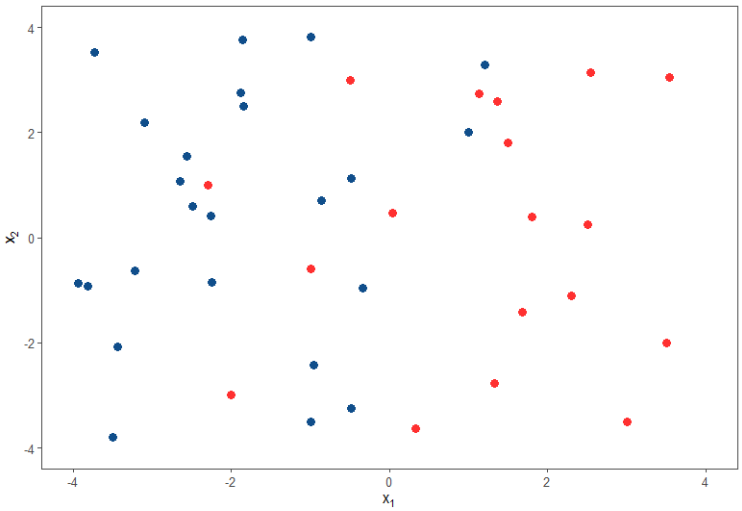

Figura 3.5: Os pontos em azul representam os clientes adimplentes e os pontos em vermelho os clientes inadimplentes. Nesse caso, as duas classes não são linearmente separáveis, ou seja, não conseguimos encontrar um hiperplano $H_{\boldsymbol{\beta}}$, tal que as duas classes sejam separadas perfeitamente. Dessa forma, não é possível aplicar o classificador de margem máxima.

O classificador de vetores suporte é um caso mais geral do classificador de margens máximas, pois conseguimos utilizá-lo em uma diversidade maior de conjunto de dados, pelo fato de ser mais flexível quanto a suposição de separação linear do conjunto de dados. Tal classificador também tem como característica uma maior robustez na classificação, pois, conforme abordado anteriormente, a definição de um hiperplano ótimo depende apenas dos vetores suporte, e uma pequena alteração na posição desses vetores pode resultar em um hiperplano totalmente diferente. O classificador de vetores suporte, no

entanto, é baseado em um hiperplano que não separa perfeitamente as duas classes a fim de garantir melhor classificação dos indivíduos ou objetos do conjunto de treinamento e, consequentemente, mais flexibilidade na classificação de uma nova observação. Nesse sentido, o classificador de vetores suporte permite que uma observação seja classificada do lado errado da margem ou até mesmo do hiperplano.

Nesse caso, buscamos o hiperplano que seja solução do seguinte problema de otimização:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1},\boldsymbol{\epsilon}\in\mathbb{R}^{n+1}}{\text{maximizar}} \; M\left(H_{\boldsymbol{\beta}}\right) \tag{3.17}$$

$$\text{Sujeito a } \sum_{j=1}^{p}\beta_j^2 = 1, \tag{3.18}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M\left(H_{\boldsymbol{\beta}}\right)(1-\epsilon_i), \tag{3.19}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n}\epsilon_i \leq c, \tag{3.20}$$

em que c é uma parâmetro real de ajuste não-negativo, $M(H_{\boldsymbol{\beta}})$ é a margem associada ao hiperplano $H_{\boldsymbol{\beta}}$ e $\epsilon_1, \ldots \epsilon_n$ são variáveis de "folga" que permitem observações individuais serem classificadas do lado errado da margem ou até mesmo do hiperplano. Nesse caso, queremos encontrar o vetor de parâmetros $\boldsymbol{\beta}$ que resulte no hiperplano que satisfaz as condições de (3.18) - (3.20). Uma vez que tal hiperplano é encontrado, conseguimos classificar uma nova observação $\boldsymbol{x}^* \in \mathbb{R}^p$, avaliando o sinal da função linear $f : \mathbb{R}^p \to \mathbb{R}$ tal que

$$f\left(\boldsymbol{x}^*\right) = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*.$$

As condições do problema de otimização para o classificador de vetores suporte se assemelham às condições do problema de otimização visto anteriormente, do classificador de margem máxima. Todavia, nesse caso, temos novos argumentos que possibilitam a classificação de uma observação do lado errado da margem ou do hiperplano: as variáveis de "folga" e o parâmetro de ajuste. Para todo número natural i tal que $1 \leq i \leq n$, a variável ϵ_i indica onde a i-ésima observação está localizada relativa ao hiperplano e relativa à margem. Se $\epsilon_i = 0$, a i-ésima observação está classificada do lado correto da margem; se $0 < \epsilon_i < 1$, a i-ésima observação violou o limite da margem; e se $\epsilon_i \geq 1$, a i-ésima observação está localizada do lado errado do hiperplano. A constante c é considerada um

hiperparâmetro que limita a soma dos erros, ou seja, é o valor que controla o número de violação e a gravidade dessas violações. Quando $c = 0$, temos $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_n = 0$, então não é tolerado que nenhuma observação viole a margem e a condição (3.19) se iguala a (3.13). Quando $c > 0$, temos no máximo c observações do lado errado do hiperplano. Assim, conforme o valor de c aumenta, nos tornamos mais tolerantes com o número de observações que podem violar a margem. Por outro lado, quando o valor de c diminui nos tornamos menos tolerantes com relação a essa violação.

Na Figura 3.6, abordamos um exemplo da obtenção do hiperplano via classificador de vetores suporte, o qual admite que algumas observações sejam classificadas além da margem ou até mesmo do lado errado do hiperplano.

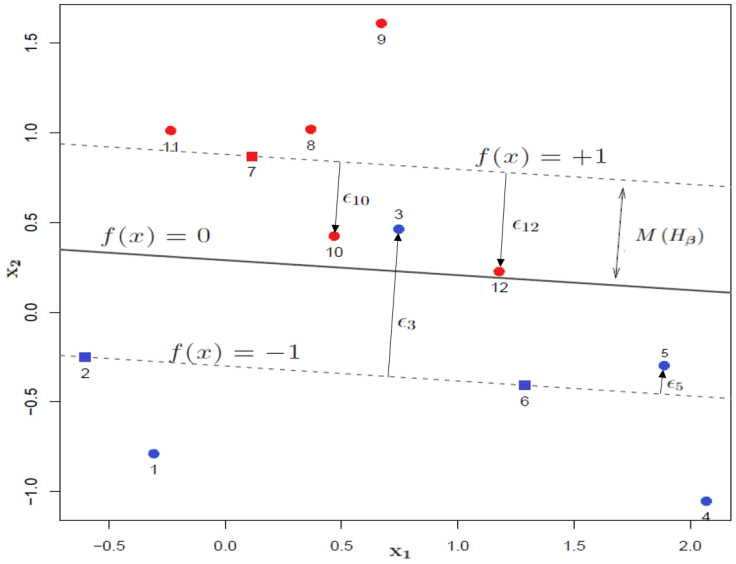

Figura 3.6: Os pontos em azul simbolizam os clientes adimplentes e os pontos em vermelho os inadimplentes. As observações representadas por um "quadrado"são os vetores suporte que orientam a obtenção do hiperplano. As observações 5, 10 e 12 estão violando o limite de sua respectiva margem, logo são observações classificadas erroneamente, de modo que $\epsilon_5, \epsilon_{10}, \epsilon_{12}$ são valores entre 0 e 1 . Para a observação 3, $\epsilon_3 > 1$ dado que está classificada do lado errado do hiperplano.

Analisando o problema de otimização (3.17) - (3.20), notamos que somente as observações que compõem a margem ou violam os limites da margem ou do hiperplano é que são determinantes para obtenção do hiperplano. Esses pontos são denominados de vetores suporte. O parâmetro c também é determinante na obtenção do classificador, pois para valores grandes de c, estamos mais tolerantes a violações e, consequentemente, teremos margens maiores; por outro lado, se c for um número pequeno, poucas violações serão toleradas e as margens tendem a ser menores.

3.1.6 Máquinas de vetores suporte

Como alternativa para lidar com conjunto de dados que não são linearmente separáveis, podemos utilizar o classificador de vetores suporte introduzido na seção anterior, o qual encontra um limite linear, também denominado de hiperplano, para realizar a classificação de novas unidades amostrais de modo a flexibilizar a classificação de algumas observações do lado errado da margem ou até mesmo do hiperplano. Para encontrar tal hiperplano, utilizamos o espaço das covariáveis de entrada, ou seja, o hiperplano de separação é obtido utilizando as variáveis explicativas em sua dimensão original. Entretanto, na grande maioria dos conjuntos de dados utilizados em estudos reais, não conseguimos encontrar um limite linear mesmo aplicando a flexibilização proposta pelo classificador de vetores suporte, pois as classes apresentam um relacionamento mais complexo.

Nessa situação, para generalizar a aplicação do classificador de vetores suporte, aplicamos uma transformação não-linear arbitrária Φ no conjunto de treinamento, buscando a projeção do espaço original p-dimensional em um espaço de maior dimensionalidade, denominado de espaço de características. Dessa forma, teremos maiores chances de obter dados linearmente separáveis. Tal afirmação é baseada no Teorema de Cover (Cover, 1965), em que segundo o autor um conjunto de dados não linearmente separável tem maior probabilidade de ser linearmente separável no espaço de características.

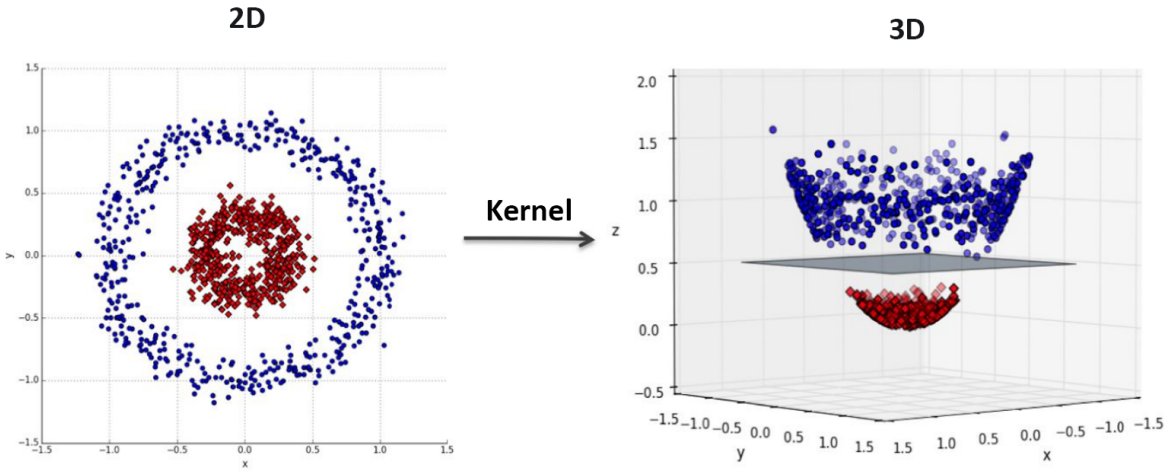

Figura 3.7: No plano cartesiano à esquerda temos a representação de um conjunto de treinamento bidimensional não linear. Após a aplicação da função Φ (futuramente definiremos como *Kernel*) obtemos um novo conjunto de treinamento com dimensionalidade três, representado pelo plano cartesiano à direita, no qual torna-se linear no espaço de características. **Fonte:** Hachimi *et al.* (2020).

A Figura 3.7 ilustra um exemplo de como a aplicação da função Φ no conjunto de treinamento pode tornar os dados linearmente separáveis no espaço de características, facilitando a obtenção do hiperplano ótimo.

Sejam p e q números naturais não-nulos tais que $p < q$. Considere conjunto de treinamento $Z = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, em que $\boldsymbol{x}_i \in \mathbb{R}^p$ é o vetor de características observadas associado ao indivíduo i e $y_i \in \{1, -1\}$ é a classificação do indivíduo i em adimplente ou inadimplente, em que i e n são números naturais tais que $1 \leq i \leq n$. Definimos a função não linear $\Phi : \mathbb{R}^p \to \mathbb{R}^q$, que mapeia o vetor de características observadas $\boldsymbol{x_i} \in \mathbb{R}^p$ associado ao indivíduo i a um novo vetor de características $\phi(\boldsymbol{x}_i) \in \mathbb{R}^q$, $\forall i = 1, 2, \ldots, n$. Isso nos leva a um novo conjunto de treinamento $Z^{'} = \{(\Phi(\boldsymbol{x}_1), y_1), \ldots, (\Phi(\boldsymbol{x}_n), y_n)\}$. Podemos mostrar que a obtenção do classificador de vetores suporte a partir desse novo conjunto de treinamento $Z^{'}$ se resume ao mesmo problema de otimização abordado em (3.17) - (3.20), no qual devemos apenas substituir o conjunto de treinamento, aplicando o conjunto mapeado no espaço de características. Em outras palavras,

$$\underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^q, \boldsymbol{\epsilon} \in \mathbb{R}^n}{\text{maximizar}} \; M\left(H_{\boldsymbol{\beta}}\right) \tag{3.21}$$

$$\text{Sujeito a } \sum_{i=1}^{q} \beta_i^2 = 1, \tag{3.22}$$

$$y_i(\beta_0 + \boldsymbol{\beta}^\top \Phi(\boldsymbol{x_i})) \geq M\left(H_{\boldsymbol{\beta}}\right)(1 - \epsilon_i), \tag{3.23}$$

$$\epsilon_i \geq 0, \; \sum_{i=1}^{n} \epsilon_i \leq c. \tag{3.24}$$

Note que existem muitas possibilidades para ampliar a dimensionalidade do espaço das variáveis originais com a escolha da função Φ, dessa forma, é fácil obter um mapeamento com muitas variáveis, ou até um cujo espaço das covariáveis é de dimensão infinita, o que tornaria a obtenção do classificador muito custosa computacionalmente. Nesse contexto, foi proposto o classificador de máquinas de vetores suporte, uma extensão do classificador de vetores suporte, que amplia a dimensão do espaço de variáveis de forma específica, utilizando *kernels*. Os *kernels* são funções que quantificam a similaridade de duas observações, e tendem a ser uma maneira mais prática e eficaz para obtenção de um classificador.

De acordo com Izbicki e dos Santos (2020) se $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ é a solução do problema de otimização abordado em (3.17) - (3.20) podemos reescrever o classificador $f : \mathbb{R}^p \to \mathbb{R}$ associado ao hiperplano ótimo $H_{\boldsymbol{\beta}}$ da seguinte maneira:

$$f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle \boldsymbol{x}, \boldsymbol{x_i} \rangle, \tag{3.25}$$

em que $\langle \boldsymbol{x}, \boldsymbol{x_i} \rangle$ é o produto interno de dois vetores definido como $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{k=1}^{p} a_k b_k$; e α_i é um parâmetro associado a i-ésima observação do conjunto de treinamento. Note que existe uma correspondência entre o parâmetro α_i e os parâmetros originais β_j. A partir de (3.25) notamos que para encontrar o classificador de vetores suporte e estimar os parâmetros necessitamos apenas do produto interno entre todas as observações. Além disso, podemos mostrar que $\alpha_i = 0$ para todos os pontos que não forem vetores suporte (Gareth *et al.*, 2013). Assim, seja S um conjunto que contêm os índices dos vetores suporte, podemos reescrever (3.25) da seguinte maneira:

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i \in S} \alpha_i \langle \boldsymbol{x}, \boldsymbol{x_i} \rangle. \tag{3.26}$$

De forma análoga, segue que a solução $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ do problema de otimização, com o espaço das covariáveis ampliado, abordado em (3.21) - (3.24), nos leva a um classificador $f : \mathbb{R}^p \to \mathbb{R}$ dado por

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i \in S} \alpha_i \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x_i}) \rangle. \tag{3.27}$$

Nesse caso, poderíamos ter dificuldade de calcular o produto interno $\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x_i}) \rangle$, pois além da necessidade de definir a função Φ, a transformação resultante pode gerar um grande número de variáveis, ou até mesmo infinitas variáveis no espaço de características, o que tornaria a obtenção do classificador um processo complexo e custoso. Nesse sentindo, a proposta dos classificadores de máquinas de vetores suporte é a substituição do produto interno $\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x_i}) \rangle$, por um *kernel* genérico $\kappa(\boldsymbol{x}, \boldsymbol{x_i})$, resultando no classificador $f : \mathbb{R}^p \to \mathbb{R}$ tal que

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i \in S} \alpha_i \kappa(\boldsymbol{x}, \boldsymbol{x_i}). \tag{3.28}$$

Note que, utilizando *kernels* não é necessário calcular o produto interno, o que torna o processo de obtenção do classificador de máquinas de vetores suporte mais eficiente e menos custosa. O cálculo do produto interno $\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x_i}) \rangle$ não se faz necessário, pois de

acordo com o truque do *kernel* (ver página 65, Teorema 4, de Izbicki e dos Santos (2020)), qualquer *kernel* de Mercer pode ser representado por:

$$\kappa(\boldsymbol{x}, \boldsymbol{x_i}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x_i}) \rangle.$$

Definição 3.29 (Kernel de Mercer) *Seja p um número natural não nulo e $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ vetores das características observadas, em que $\boldsymbol{x_i} \in \mathbb{R}^p$ é o vetor de características observadas associado ao indivíduo i, em que i é um número natural tal que $1 \leq i \leq n$. Uma função $\kappa : \mathbb{R}^p \to \mathbb{R}$ é dita ser um **Kernel de Mercer** se possuir as seguintes propriedades:*

- ***Simetria**: $\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \kappa(\boldsymbol{x_j}, \boldsymbol{x_i})$, para todo $\boldsymbol{x_i}, \boldsymbol{x_j} \in \mathbb{R}^p$;*
- ***Positiva semidefinida**: a matriz K de ordem $n \times n$ tal que o elemento genérico k_{ij} que está na linha i e coluna j da matriz K é tal que $k_{ij} = \kappa(\boldsymbol{x_i}, \boldsymbol{x_j})$ e a matriz K é positiva semidefinida.*

Exemplo 3.30 *Os kernels de Mercer mais utilizados são:*

1. ***Polinomial:***

$$\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = (1 + \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle)^d.$$

O expoente natural d deve ser definido a priori antes de calcular o classificador pelo usuário.

2. ***Gaussiano (ou RBF):***

$$\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \mathbf{exp}\left(\frac{||\boldsymbol{x_i} - \boldsymbol{x_j}||^2}{2\sigma^2}\right).$$

O parâmetro real que controla a amplitude $\sigma^2 > 0$, comum a todos os Kernels, deve ser especificado previamente.

3. ***Sigmoidal:***

$$\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \mathbf{tanh}(\beta_0 \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle) + \beta_1$$

Vale ressaltar, que no Kernel sigmoidal, a depender do valor de β_0 e β_1 , o teorema de Mercer não é satisfeito. Nesse sentido, a utilização do Kernel sigmoidal fica restrita para alguns valores desses parâmetros.

Logo, para obtermos o classificador de máquinas de vetores suporte, devemos realizar algumas escolhas como, a função *kernel* a ser utilizado e seus respectivos parâmetros, bem como o valor do parâmetro c. (Gonçalves, 2015).

3.1.7 Processo de otimização

O objetivo principal desta monografia é abordar – de forma geral – o funcionamento do classificador de máquinas de vetores suporte. Nesse sentido, não nos aprofundamos, matematicamente, no processo de obtenção dos resultados dos problemas de otimização proposto na Seção 3.1.5. Entretanto, para entendimento de assuntos abordados nas seções subsequentes, foi necessário explicitar alguns passos que ocorrem no processo de otimização para obtenção do hiperplano ótimo.

De acordo com Hastie *et al.* (2009) e Ng (2000) o problema de otimização definido em (3.17) - (3.20), para obtenção do hiperplano de margens flexíveis, pode ser reescrito da seguinte forma

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\beta_0}}{\text{minimizar}} \left(\frac{1}{2}\left\|\boldsymbol{\beta}\right\|^2 + C\sum_{i=1}^{n}\epsilon_i\right) \tag{3.31}$$

$$\text{Sujeito a } \epsilon_i \geq 0,\ y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq (1-\epsilon_i), \quad i = 1,\ldots,n, \tag{3.32}$$

em que $\left\|\boldsymbol{\beta}\right\|^2 = \sum_{j=1}^{p}\beta_j^2$, ϵ_i variáveis de folga, que medem o quanto uma observação está mal classificada e C pode ser definido como um parâmetro de custo. Note, que nesse caso, a constante C tem interpretação inversa aquela apresentada na Equação (3.20), em que pela definição, c tem a função de limitar a soma do erro de má classificação. No problema de otimização Equação (3.31), a constante C que acompanha o erro ϵ_i tem a finalidade de ser como um custo de má classificação, além disso, pode ser entendida como $trade-off$ entre atingir um baixo erro nos dados de treinamento e maximizar a margem.

Notamos, que dessa forma, a restrição $\sum_{j=1}^{p}\beta_j^2 = 1$ ou $\left\|\boldsymbol{\beta}\right\|^2 = 1$, que era um empecilho para aplicação de métodos tradicionais de otimização por conta de ser não convexa, não se faz mais necessária. Sendo assim, agora com a função objetivo e as restrições convexas temos um problema de otimização convexa, que pode ser resolvido pelo método clássico de programação quadrática: os multiplicadores de Lagrange.

Seguindo o exposto por Ng (2000) a função primal de Lagrange, a qual queremos

minimizar, é dada por:

$$\mathcal{L}_{(\boldsymbol{\beta},\beta_0,\epsilon,\alpha,r)} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}\epsilon_i - \sum_{i-1}^{n}\alpha_i\left[y_i\left(\boldsymbol{x_i^t\beta} + \beta_0\right) - (1-\epsilon_i)\right] - \sum_{i=1}^{n} r_i\epsilon_i, \tag{3.33}$$

em que α_i e r_i são os multiplicadores de Lagrange. Seguindo com a solução da Equação (3.33), derivando os parâmetros que queremos minimizar obtemos a seguinte forma dual lagrangeana:

$$\max_{\alpha}\left\{\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle \boldsymbol{x_i}, \boldsymbol{x_j}\right\rangle\right\} \tag{3.34}$$

$$\text{Sujeito a}\sum_{i=1}^{n} y_i\alpha_i = 0, \tag{3.35}$$

$$0 \leq \alpha_i \leq C \quad i = 1,\ldots,n. \tag{3.36}$$

Além disso, as seguintes condições de Karush-Kuhn-Tucker (KKT) devem ser satisfeitas (Hastie *et al.*, 2009):

$$\alpha_i\left[y_i\left(\boldsymbol{x_i}^t\boldsymbol{\beta} + \beta_0\right) - (1-\epsilon_i)\right] = 0, \tag{3.37}$$

$$r_i\epsilon_i = 0, \tag{3.38}$$

$$y_i\left(\boldsymbol{x_i}^t\boldsymbol{\beta} + \beta_0\right) - (1-\epsilon_i) \geq 0. \tag{3.39}$$

Ao final do processo, encontramos os valores dos parâmetros $\boldsymbol{\beta}$, β_0 e α_i que determinam a equação do hiperplano ótimo. Essa abordagem, também pode ser aplicada no problema de otimização para obtenção do hiperplano ótimo, no caso de dados separavéis, e também no caso mais geral, quando utilizamos os *Kernels*.

3.1.8 Máquinas de vetores suporte e dados desbalanceados

Os classificadores, de modo geral, tendem a apresentar baixo desempenho na classificação de novas observações quando o conjunto de dados apresenta um significativo desequilíbrio entre a proporção das classes. O mesmo comportamento pode ser observado para o classificador de máquinas de vetores suporte, no qual em um cenário de desbalanceamento de classes, o hiperplano de separação pode ser enviesado para classe minoritária, ou seja, o limite de decisão é "atraído" para próximo das observações pertencentes a essa classe, o que compromete o desempenho do classificador (Batuwita e Palade, 2013).

De acordo com Akbani *et al.* (2004) e Tang *et al.* (2008), tal viés pode ser ocasionado em decorrência do processo de otimização de margem flexível, abordado na Equação (3.31). Nesse processo, queremos encontrar o vetor de parâmetros $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ que maximize a margem $M(H_{\boldsymbol{\beta}})$, que é equivalente a minimizar $\frac{1}{2}\|\boldsymbol{\beta}\|^2$, e minimize o termo de penalização dado por $\sum_i^n \epsilon_i$. Em tal contexto, o parâmetro C funciona como um *trade-off*, em que sua escolha determina se estamos mais disposto a maximizar a margem ou minimizar os erros. Note, que esse parâmetro é refletido no erro de modo geral, o que implica que o custo de má classificação é igual tanto para classe minoritária quanto para classe majoritária, desse modo, na busca por minimizar o termo de penalidade, a região de aceitação da classe majoritária é aumentada, empurrando o hiperplano para próximo da classe minoritária, pois assim, o erro acumulativo na classe com maior densidade de observações (classe majoritária) tende a zero. Em contrapartida, o erro aumenta na classe minoritária, porém como a frequência de observações é baixa, o impacto gerado no termo de penalidade é mínimo. Em casos de extremo desequilíbrio entre as classes, o classificador de máquinas de vetores suporte pode gerar hiperplanos amplamente distorcidos que classifica todas as novas observações na classe majoritária (no contexto desse estudo, classifica como clientes adimplentes). Tal viés, aumenta consideravelmente a previsão de falsos negativos no conjunto de teste.

Outro fator que pode contribuir com o baixo desempenho do classificador de máquinas de vetores suporte, em um contexto de desbalanceamento de classes, é o desequilíbrio da proporção de vetores suporte. Em seus estudos Wu e Chang (2003) observaram, experimentalmente, que o aumento do desequilíbrio entre as classes, na amostra de treinamento, gera um aumento da desproporcionalidade dos vetores suporte, de modo que a densidade de vetores suporte da classe majoritária é superior a dos vetores suporte da classe minoritária. Nesse sentido, os autores relatam que uma amostra de teste situada perto da fronteira de separação (hiperplano) tem muito mais chances de ficar rodeada pelo vetores suporte da classe majoritária, o que aumenta a probabilidade dessas observações serem classificadas em tal classe. Entretanto, Akbani *et al.* (2004) argumentam que tal apontamento pode ser amenizado, em níveis moderados de desbalanceado de classes, por conta da restrição imposta na Equação (3.35), em que a soma dos vetores suportes ponderados pelos α_i, que funcionam como pesos para função de decisão (Equação (3.28)), devem ser iguais para as duas classes. Isto posto, para a condição ser satisfeita, os α_i da classe minoritária devem ser maiores que o da classe majoritária, o que implica que os vetores suporte

da classe com menos obervações tendem a ter mais influência na decisão do classificador para compensar a superioridade de vetores suporte da classe majoritária. Na Figura 3.8 abordamos um exemplo do deslocamento que o hiperplano sofre quando ajustados com conjunto de dados desbalanceado.

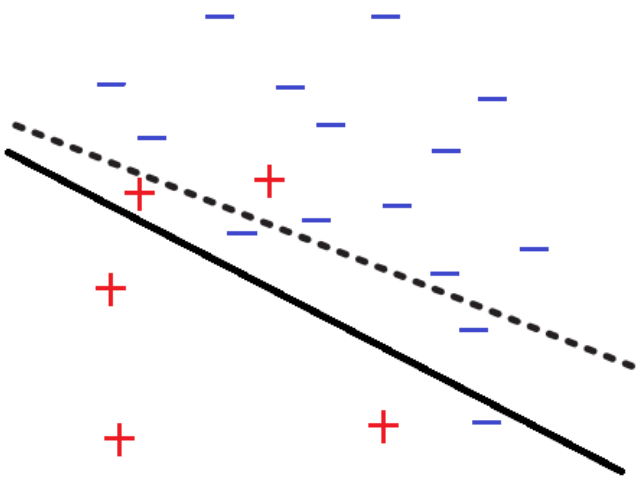

Figura 3.8: Representação do hiperplano ajustado em um cenário de desequilíbrio de classes. Os pontos representados por " - "em azul simbolizam a classe majoritária (clientes adimplentes) enquanto que os pontos representados por " + " simbolizam a classe minoritária (clientes inadimplentes). A linha em preto pontilhada simboliza o limite ideal para separação e a linha em preto preenchida representa o hiperplano obtido com a aplicação do classificador de máquina de vetores suporte. **Fonte:** Adaptado de Phoungphol *et al.* (2012)

Desse modo, na tentativa de melhorar o desempenho do classificador de máquinas de vetores suporte com conjunto de dados desbalanceados, estudamos algumas metodologias propostas na literatura. Basicamente, podemos seguir duas abordagens, aplicar um pré-processamento no conjunto de dados buscando equilibrar as classes (denominado de métodos externos) ou aplicar modificações no algoritmo de classificação (denominado de métodos internos). Nessa monografia, utilizamos o SMOTE (*Synthetic Minority Oversampling Technique*) como método de pré-processamento dos dados e SVM sensível ao custo como método interno.

3.1.9 Máquinas de vetores suporte sensível ao custo

Conforme dissertado anteriormente, uma das causas para o baixo desempenho do classificador de máquinas de vetores suporte em um cenário de desequilíbrio de classe, é que na busca de diminuir o termo de penalidade o hiperplano é enviesado para classe minoritária. Umas das suposições para ocorrência de tal efeito é que o custo é fixo para

ambas as classes. Nesse sentido, Veropoulos *et al.* (1999) propôs em seu estudo utilizar diferentes custos de má classificação para as classes, de forma que a classificação incorreta na classe minoritária tenha maior penalização do que na classe majoritária. Basicamente, para tal realização a formulação primal lagrangiana, dada pela Equação (3.33) sofre a seguinte modificação:

$$\mathcal{L}_{(\boldsymbol{\beta},\beta_0,\epsilon,\alpha,r)} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C^+ \sum_{\{i|y_i=+1\}}^{n} \epsilon_i + C^- \sum_{\{i|y_i=-1\}}^{n} \epsilon_i - \sum_{i-1}^{n} \alpha_i \left[y_i\left(\boldsymbol{x_i^t}\boldsymbol{\beta} + \beta_0\right) - (1-\epsilon_i)\right] - \sum_{i=1}^{n} r_i\epsilon_i,$$

em que $\alpha_i \geq 0$ e $r_i \geq 0$. Note que denotamos por C^+ o custo de classificação incorreta para classe minoritária e C^- o custo de classificação incorreta para classe majoritária.

É direto mostrar que a forma dual lagrangiana segue a mesma que a exposta na Equação (3.34), porém com as seguintes restrições para α_i: se $y_i = +1$, então

$$0 \leq \alpha_i \leq C^+$$

e se $y_i = -1$, temos

$$0 \leq \alpha_i \leq C^-.$$

Em seu estudo, Veropoulos *et al.* (1999) não apresentaram a forma como devem ser definidos os valores de C^+ e C^-. Entretanto, em alguns estudos presentes na literatura os autores sugerem fixar o valor do custo para classe majoritária, por exemplo $C^- = 1$, e valor do custo da classe minoritária é dado pela razão do número de observações da classe majoritária pelo número de observações da classe minoritária. Além dessa opções, também podemos realizar estudos empíricos utilizando diferentes valores para parâmetros de custo e observar qual resulta na melhor performance do classificador.

3.1.10 SMOTE

Como abordado nessa monografia, utilizar um conjunto de dados com classes desbalanceadas pode comprometer consideravelmente o desempenho de predição dos classificadores, em que a chance de ocorrer falsos negativos aumentam consideravelmente, ou seja, em

tais condições de desequilíbrio, os classificadores tendem a classificar as novas observações com maior frequência na classe majoritária, pois esse seria o caminho mais simples para o algoritmo (Akbani *et al.*, 2004). Esse tipo de inflação no erro de classificação pode ser muito custoso em determinadas áreas como por exemplo, na disponibilização de crédito ou na medicina, com o diagnóstico de doenças raras.

Para mitigar tal efeito, é comum realizar uma reamostragem no conjunto de treinamento de forma a sobreamostrar a classe minoritária ou subamostrar a classe majoritária, a fim de equilibrar a distribuição das classes. Entretanto, esse tipo de procedimento pode acarretar no enviesamento do novo conjunto de treinamento, pois realizando uma subamostragem da classe majoritária estaremos perdendo informação de observações que podem ser relevantes, e por outro lado, sobreamostrando estaremos duplicando essas informações. Além disso, podemos combinar os métodos de subamostragem com os métodos de sobreamostragem. No entanto, estudos relataram que esses métodos híbridos não levam a um melhor desempenho na classificação de novas unidades amostrais (Chawla *et al.*, 2002).

Na busca por encontrar uma nova solução para o problema exposto, Chawla *et al.* (2002) propuseram uma técnica denominada SMOTE (*Synthetic Minority Oversampling Technique*) - técnica de sobreamostragem minoritária sintética - que tem como objetivo inflar a classe minoritária, criando indivíduos sintéticos com características muito semelhante aqueles já pertencentes ao conjunto de dados inicial. Dessa forma, estamos incluindo indivíduos que não são exatamente uma cópia daqueles já existente na base, mas que tem um comportamento muito próximo do esperado caso fosse possível amostrar mais indivíduos daquela classe.

O processo de implementação consiste em fazer uma interpolação linear entre os K vizinhos mais próximos de uma observação da classe minoritária amostrada aleatoriamente, e adicionar um valor aleatório entre 0 e 1 ao vetor de características dos K vizinhos mais próximo da observação selecionada. De forma detalhada, o algoritmo SMOTE pode ser definido de acordo com os seguintes procedimentos:

1. Defina N, tal que $N \in \mathbb{N}$, como sendo a proporção de sobreamostragem desejada, de forma que N multiplicado pelo número de observações da classe minoritária será a quantidade observações sintéticas resultantes do algoritmo. Por exemplo, suponhamos que temos 1000 observações na classe minoritária e gostaríamos que ao final do processo ficássemos com 5000 observações. Então, seria necessário criar

4000 observações sintéticas, um número 4 vezes maior que o tamanho original da classe. Logo, deveríamos definir $N = 4$;

2. Defina o valor K, $K \in \mathbb{N}$, como sendo a quantidade de vizinhos mais próximos que será utilizada pelo algoritmo;

3. Inicia-se o processo iterativo para criação das unidades sintéticas, da seguinte forma:

   (a) Seleciona-se aleatoriamente uma observação $\boldsymbol{x}_i$ pertencente a classe minoritária;

   (b) Encontra-se os K vizinhos mais próximos de $\boldsymbol{x}_i$, a partir de uma medida de distância, por exemplo a distância euclidiana.

   (c) Seleciona-se aleatoriamente, com repetição, N observações do conjunto dos K vizinhos mais próximos.

   (d) Em seguida calcula-se a diferença entre o vetor de características da observação selecionada e cada um dos N vizinhos selecionados;

   (e) Por fim, para cada N observação selecionadas no item (c) gera-se uma nova observação sintética somando-se ao vetor de característica da observação selecionada a diferença, obtida no passo anterior, multiplicada por um valor entre 0 e 1, gerado de uma distribuição uniforme. A definição de uma nova observação sintética $\boldsymbol{s}_{ij}$ é dada pela seguinte expressão:

$$\boldsymbol{s}_{ij} = \boldsymbol{x}_i + unif(0,1) \times (\boldsymbol{x}_{ij} - \boldsymbol{x}_i) \qquad i = 1, \ldots, n \text{ e } j = 1, \ldots, N.$$

4. Repita o processo selecionando outras obervações do conjunto minoritário até que a quantidade de amostras geradas seja equivalente a multiplicação de N pelo número de linhas da classe minoritária.

Dessa forma, ao final do algoritmo iremos obter uma nova amostra da classe minoritária, composta pelas observações, originalmente pertencentes à classe, e as novas observações sintéticas geradas pelo método descrito.

3.2 Medidas de performance

O principal objetivo dessa monografia é avaliar a performance do classificador de máquinas de vetores suporte (SVM) em um conjunto de dados com classes desbalancea-

das, aplicando o método tradicional e o método do SVM sensível ao custo, assim como avaliar o desempenho do classificador no mesmo conjunto de dados após aplicação do método de pré-processamento de dados SMOTE. Para avaliar tal performance, utilizaremos como principal ferramenta as medidas obtidas da matriz de confusão, a partir da qual podemos definir métricas de performance como acurácia, sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo, coeficiente de correlação de Matthews e a GMédia. Para fins de comparação também analisamos o desempenho de outros classificadores como regressão logística e florestas aleatórias.

3.2.1 Matriz de confusão

A matriz de confusão poder ser vista como uma tabela ou matriz que resume os resultados da predição de um modelo de classificação, fornecendo informações acerca dos erros e acertos cometidos pelo classificador. Em geral, ela é utilizada com modelos cujo banco de dados contém apenas duas classes, entretanto, também pode ser aplicada em conjunto de dados com três ou mais classes. No caso binário (duas classes), a matriz é construída com 2 colunas e 2 linhas, em que nas linhas representamos os valores observados das classes e nas colunas os valores previstos, ou vice e versa. Um exemplo de matriz de confusão para resultado de um modelo de classificação de clientes em adimplentes e inadimplentes pode ser visualizado na Tabela 3.1.

Tabela 3.1: Matriz de confusão para classificação binária no contexto de análise de crédito: n é o número de clientes no conjunto de teste, i_I é o número de clientes inadimplentes classificados corretamente, i_A é o número de clientes inadimplentes classificados incorretamente, a_I é o número de clientes adimplentes classificados incorretamente, a_A é o número de clientes adimplentes classificados corretamente, i é o número de clientes inadimplentes no conjunto de teste, a é o número de clientes adimplentes no conjunto de teste, I é o número de clientes classificados como inadimplentes e A é o número de clientes classificados como adimplentes.

Valores previstos	Valores observados		
pelo classificador	inadimplente	adimplente	Total
inadimplente	i_I	a_I	I
adimplente	i_A	a_A	A
Total	i	a	n

Preenchida a tabela com os valores observados na amostra e com o resultado da classificação dos modelos, podemos utilizar essas informações para construir métricas que irão auxiliar na análise da performance do classificador.

3.2.2 Medidas de performance baseadas na matriz de confusão

Em concordância com a nomenclatura, comumente apresentada na literatura, com a Tabela 3.1 e definindo a classe Inadimplente como a classe positiva, denotamos por

- VP o número de verdadeiros positivos, isto é, a quantidade de clientes inadimplentes classificados como inadimplentes ($VP = i_I$);
- VN é o número de verdadeiros negativos, isto é, a quantidade de clientes adimplentes classificados como adimplentes ($VN = a_A$);
- FP é o número de falsos positivos, isto é, a quantidade de clientes adimplentes classificados como inadimplentes ($FP = a_I$);
- FN é o número de falsos negativos, isto é, a quantidade de clientes inadimplentes classificados como adimplentes. ($FN = i_A$).

Baseado nessas definições, podemos calcular as seguintes métricas:

- **Acurácia:** é o número AC que quantifica a proporção de clientes que foram classificados corretamente, isto é,

$$AC = \frac{VP + VN}{n} = \frac{i_I + a_A}{n}.$$

  É importante destacar, a necessidade de ser cauteloso ao analisar somente essa medida, pois em alguns casos, quando o conjunto de dados apresenta algum grau de desequilíbrio, os classificadores tendem a classificar a maioria das observações na classe majoritária, o que acarreta em uma grande quantidade de acerto da classe mais frequente, embora erre na classe de menor frequência. Nesse caso, apesar do baixo desempenho na predição da classe minoritária o valor de AC pode ser elevado.

- **Taxa de erro:** é o número TE que quantifica a proporção de clientes que foram classificados incorretamente, isto é,

$$TE = \frac{FP + FN}{n} = \frac{a_I + i_A}{n} = 1 - AC.$$

  Note que essa medida é complementar à medida de acurácia.

- **Sensibilidade:** é o número S que quantifica a proporção de clientes inadimplentes que foram classificados corretamente, isto é,

$$S = \frac{VP}{VP + FN} = \frac{i_I}{i_I + i_A} = \frac{i_I}{i}.$$

  Em algumas referências podemos encontrar a sensibilidade denominada como *Recall*.

- **Especificidade:** é o número E que quantifica a proporção dos clientes adimplentes que foram classificados corretamente, isto é,

$$E = \frac{VN}{VN + FP} = \frac{a_A}{a_A + a_i} = \frac{a_A}{a}.$$

- **Valor preditivo positivo:** é o número VPP que quantifica a proporção de clientes classificados como inadimplentes que foram classificados corretamente, isto é,

$$VPP = \frac{VP}{VP + FP} = \frac{i_I}{i_I + a_I} = \frac{i_I}{I}.$$

  Em algumas referências podemos encontrar a Valor preditivo positivo denominado como Precisão.

- **Valor Preditivo Negativo:** é o número VPN que quantifica a proporção dos clientes classificados como adimplentes que foram classificados corretamente, isto é,

$$VPN = \frac{VN}{VN + FN} = \frac{a_A}{a_A + i_A} = \frac{a_A}{A}.$$

Note que todas as medidas são referentes a uma proporção, logo seus valores variam de 0 a 1. Como estamos medindo os acertos do modelo, desejamos que essas medidas estejam próxima de 1, exceto a taxa de erro.

3.2.3 Medidas de performance baseadas em análises individuais

Anteriormente, apresentamos algumas medidas que podem ser calculadas a partir do resultado da tabela de contingência, denominada matriz de confusão. Tais medidas, com exceção da acurácia e da taxa de erro, são melhores interpretadas quando analisadas conjuntamente com seus índices complementares (Sensibilidade x Especificidade, VPP x

VPN), pois o resultado individual pode levar a uma conclusão distorcida da realidade, por exemplo, um valor de VPN próximo de 1 não é um indicativo de boa performance do classificador se o VPP for próximo de 0, pois nesse caso, o classificador estaria priorizando a classificação na classe negativa, e errando bastante na classe positiva. Em vista disso, foram desenvolvidas métricas de performance que resumissem os valores dos índices abordados anteriormente em uma única medida, que expressa o desempenho de predição do classificador estudado. Dentre tais métricas, podemos citar o F_1 *score*, $G - score$ ou GMédia, coeficiente de correlação de Matthews (MCC), diagnóstico de *Odds Ratio* (DOR), entre outros utilizados no cenário de aprendizado de máquina.

No contexto do tema dessa monografia, identificamos que ambas as classes devem ser priorizadas no processo de classificação, pois a escolha por maximizar umas delas e ignorar o desempenho da outra pode acarretar em prejuízos financeiros às instituições bancárias, no sentido que a escolha de um classificador que tenha alto desempenho na predição de um cliente adimplente, porém tenha baixo poder preditivo para distinguir um cliente inadimplente, levará o banco a conceder crédito para clientes com alta propensão de ser tornar inadimplente. Por outro lado, caso o comportamento oposto seja adotado, a instituição financeira estará adotando um perfil mais conservador, deixando de conceder crédito para clientes com baixo risco de inadimplência, o que irá impactar em seus lucros com operações de crédito. Isto posto, para análise do desempenho dos classificadores de máquinas de vetores suporte, florestas aleatórias e regressão logística, utilizamos métricas que consideram ambas as classes com igual importância, sendo elas a G-Média e o coeficiente de correlação de Matthews.

G-Média

A métrica G-Média, foi sugerida no estudo de Kubat *et al.* (1997) como uma medida de performance para avaliar a predição em conjunto de dados desbalanceados. Basicamente, calcula-se a média geométrica da sensibilidade e da especificidade, conforme explicitado na Equação (3.40)

$$\text{G-Média} = \sqrt{\text{Sensibilidade} \times \text{Especificidade}}. \tag{3.40}$$

O valor de G-Média será alto quando ambas medidas, Sensibilidade e Especificidade

forem altas ou quando a diferença entre elas for pequena. Por outro lado, o valor será baixo se a diferença entre as medidas for grande, nesse caso, o valor de G-Média será puxado para perto da menor medida. Dessa forma, o classificador será penalizado caso seu poder de discriminação para as classes seja baixo.

Coeficiente de Correlação de Matthews

O coeficiente de correlação de Matthews (MCC), proposto por Matthews (1975), é uma medida de performance que calcula a correlação entre o valor observado na amostra de teste e os valores preditos pelo classificador, a partir da utilização dos valores encontrados na matriz de confusão. Em vista disso, é uma medida que retorna valores entre -1 e +1, sendo que o valor +1 indica alta relação entre a previsão e o valor observado, isto é, o classificador está prevendo corretamente, e o valor -1 representa uma previsão inversa, ou seja, as previsões do classificador estão em total desacordo com as classes observadas. O valor intermediário 0, indica ausência de relação entre a previsão e o valor observado, de modo que a predição se assemelha a aleatória.

O MCC pode ser obtido a partir da seguinte expressão:

$$\text{MCC} = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP) \times (VP + FN) \times (TN + FP) \times (TN + FN)}}. \tag{3.41}$$

Note que a Equação (3.41) considera as 4 medidas da matriz de confusão VP,VN,FP e FN, consequentemente, o cálculo do MCC considera que as duas classes (adimplentes e inadimplentes) tem igual relevância. Além disso, só obteremos um valor alto para o coeficiente se as 4 medidas da matriz de confusão forem altas.

Capítulo 4

Aplicações em dados reais

Nessa etapa, aplicamos o classificador de máquinas de vetores suporte em conjuntos de dados reais e comparamos seus resultados com os modelos de regressão logística e floresta aleatória. Os classificadores serão avaliados em dois cenários: um com o conjunto de dados desbalanceados, ou seja, utilizando os dados originais e o outro cenário com o conjunto de dados balanceado a partir do método de sobreamostragem sintética, SMOTE. Além disso, também avaliaremos a performance do classificador de máquina de vetores suporte sensível ao custo (SVMSC). Para medir o desempenho dos classificadores empregados, utilizaremos o coeficiente de correlação de Matthews (MCC), G-Média e medidas fundadas na matriz de confusão, como sensibilidade, especificidade, valor preditivo positivo (VPP) e valor preditivo negativo (VPN). Inicialmente, abordaremos a análise de forma individual para cada conjunto de dados, e finalizaremos o capítulo sumarizando os principais resultados obtidos.

4.1 Inadimplência de clientes de cartão de crédito

O conjunto de dados aborda características individuais e de pagamento da fatura de cartão de crédito de clientes de um grande banco do Taiwan, no ano de 2005. Os dados foram extraídos de um estudo conduzido por Yeh e Lien (2009) e podem ser encontrados no repositório de dados de *machine learning* da UCI (Universidade da Califórnia, Irvine), disponível em: `http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`. O conjunto de dados contêm informações de 30.000 clientes, dos quais 6.636 (22,12%) são clientes inadimplentes e 23.364 (77,88%) são clientes adimplentes. A variável de interesse, que desejamos modelar é uma variável dicotômica

que identifica a inadimplência do pagamento (1 = Sim, 0 = Não). Além disso, temos à disposição as seguintes covariáveis:

- X1: Valor do crédito concedido: inclui tanto o crédito ao titular da conta quanto o crédito familiar (dependentes);
- X2: Gênero (1 = masculino; 2 = feminino);
- X3: Educação (1 = pós-graduação; 2 = universidade; 3 = ensino médio; 4 = outros);
- X4: Estado civil (1 = casado; 2 = solteiro; 3 = outros);
- X5: Idade (ano);
- X6–X11: Histórico de pagamentos anteriores. Acompanhamento do histórico de pagamentos mensais considerando o período de abril a setembro de 2005 da seguinte forma: X6 = status do pagamento em setembro de 2005; X7 = status do pagamento em agosto de 2005; ...; X11 = status do pagamento em abril de 2005. Em que a escala de medição para status de amortização é: -2 = Não teve consumo, ou seja, valor da fatura igual a zero; -1 = pagamento integral da fatura; 0 = pagamento do valor mínimo da fatura; 1 = atraso no pagamento por um mês; 2 = atraso no pagamento por dois meses; ...; 8 = atraso no pagamento por oito meses; 9 = atraso no pagamento por nove meses ou superior;
- X12–X17: Valor da fatura do cartão de crédito. X12 = Valor da fatura do cartão de crédito em setembro de 2005; X13 = Valor da fatura do cartão de crédito em agosto de 2005; ...; X17 = Valor da fatura do cartão de crédito em abril de 2005;
- X18–X23: Valor do pagamento anterior. X18 = valor pago em setembro de 2005; X19 = valor pago em agosto de 2005; ...;X23 = valor pago em abril de 2005.

Ressaltamos que as variáveis referentes a valores monetários são expressos em dólares taiwanês.

Visando o melhor aproveitamento das variáveis e maximização dos resultados, previamente à aplicação dos classificadores descritos nesta seção, realizamos a seleção e a criação de novas variáveis a partir daquelas já presentes na base de dados. Tal procedimento, foi embasado em experiências passadas com modelagem de dados de crédito. Dessa forma, o conjunto de dados utilizados para aplicação dos métodos propostos é composto pelas seguintes variáveis:

- Idade: Idade em anos;

- Atraso máximo: Qual foi o maior tempo de atraso que o cliente teve, considerando os seis meses de referência do estudo, isto é,

$$\max(Xi) \qquad \text{i} = 6,\ldots,11\ ;$$

- Proporção Limite x Pagamento: A razão do último valor pago pelo valor de crédito concedido, ou seja,

$$\frac{X18}{X1};$$

- Quantidade de atrasos: Quantas vezes o clientes atrasou o pagamento, considerando os seis meses de referência do estudo, ou seja,

$$\sum_{i=6}^{11} \mathbb{I}\left\{Xi \geq 0\right\},$$

em que $\mathbb{I}$ denota a função indicadora;

- Proporção de pagamento: Proporção do valor pago em relação ao valor da fatura, considerando os meses de Agosto, Julho, Junho e Maio. Tal variável, é obtida a partir da expressão:

$$\frac{\sum_{i=18}^{21} Xi}{\sum_{j=13}^{16} Xj};$$

- Proporção do limite utilizado: Quanto do limite foi utilizado considerando o gasto médio dos últimos 3 meses, ou seja,

$$\frac{\frac{X12+X13+X14}{3}}{X1}.$$

4.1.1 Análise descritiva e exploratória dos dados

Previamente à aplicação dos métodos propostos no conjunto de dados inadimplência de clientes de cartão de crédito, realizamos uma análise descritiva e exploratória dos dados, a fim de obter - preliminarmente - indícios do padrão de comportamento dos clientes adimplentes e inadimplentes. Durante tal processo, observamos, a partir dos gráficos de boxplots, a presença de alguns pontos *outliers*, isto é, observações cuja alguma(s) característica(s) destoa(m) drasticamente do comportamento da grande maioria. Logo,

para lidar com esses pontos, que podem influenciar no resultado das análises, aplicamos um tratamento que consiste em encontrar o valor que representa o quantil $98,5\%$ de cada variável e, em seguida, substituir o valor das observações que ultrapassem tal quantil, pelo valor do quantil encontrado. O mesmo tratamento foi aplicado para o limite inferior, em que nesse caso, substituímos os valores das observações cuja característica observada fosse inferior ao quantil $1,5$.

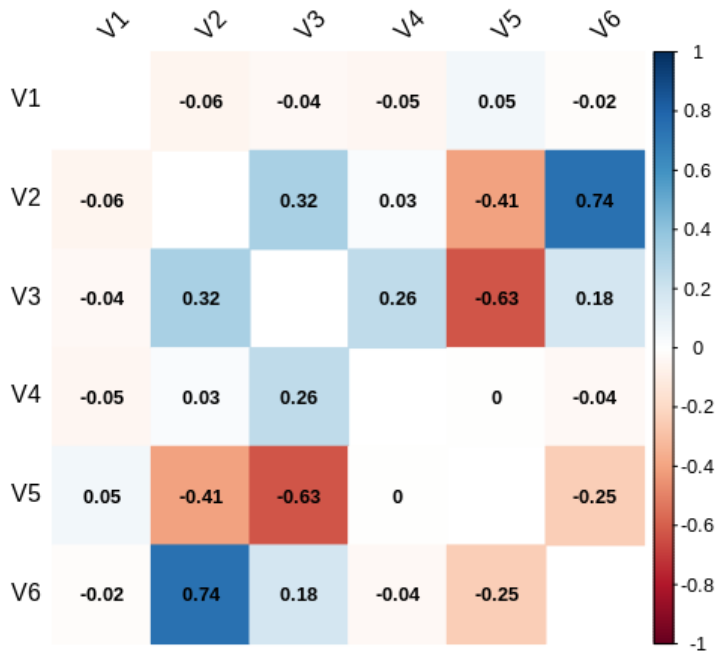

Figura 4.1: Correlograma entre as variáveis utilizadas no estudo, em que cada quadrado indica o valor e a intensidade da correlação entre as referidas variáveis. A intensidade é ilustrada a partir das cores dos quadrados, de modo que tons mais fortes de vermelho expressam forte correlação negativa, tons mais claros e próximos de branco expressam baixa correlação, sendo o branco a indicação de independência entre as variáveis, por fim tons mais escuros de azul expressam uma correlação forte e positiva. A simbologia V1,...,V6, representa de forma simplificada as seis variáveis selecionadas para prosseguimento das análises, ao passo que: V1 = Idade, V2 = Atraso máximo, V3 = Proporção do limite utilizado, V4 = Proporção limite x pagamento, V5 = Proporção de pagamento e V6 = Quantidade de atraso.

De acordo com o correlograma da Figura 4.1 observamos que as variáveis: quantidade de atraso e atraso máximo tem correlação de 0,74, o que indica uma relação positiva entre tais variáveis, de modo que clientes com maiores quantidades de atrasos tendem a atrasar por mais tempo o pagamento de suas faturas. Outra relação em destaque é entre as variáveis proporção de pagamento e proporção do limite utilizado, que apresentam uma relação negativa entre elas. Isso indica que clientes que utilizam grande parte do limite disponível, inclinam-se a pagar uma baixa proporção do valor total de sua fatura. Nessa

mesma linha, observamos que clientes que pagam uma proporção baixa do valor devido, estão suscetíveis a atrasarem o pagamento de sua fatura por mais tempo, dado que a correlação entre as variáveis de -0,41. Os demais cruzamento apresentaram baixo valor de correlação.

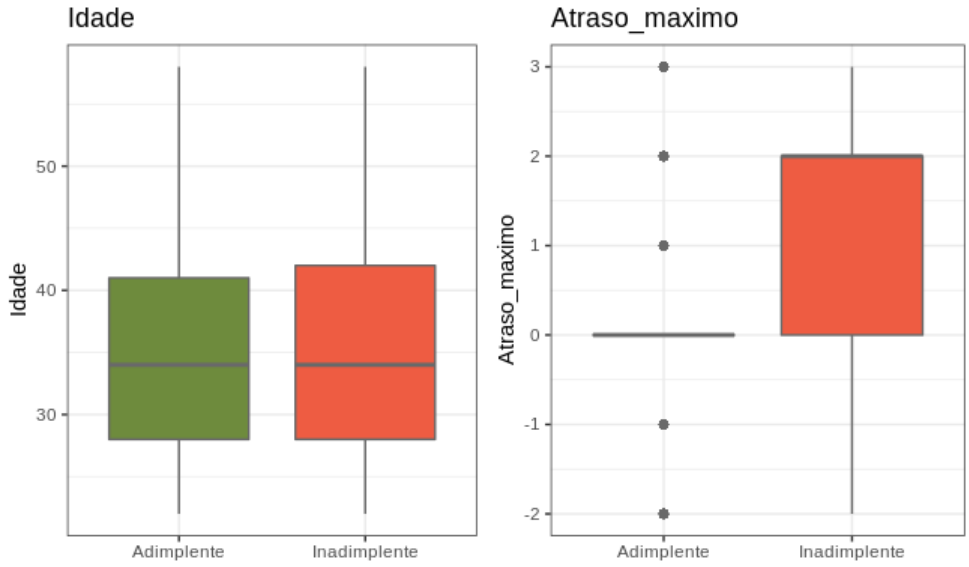

Figura 4.2: Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para Idade. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Atraso máximo.

Analisando os boxplots das variáveis ilustrados na Figura 4.2, notamos que a idade não é um fator de discriminação para inadimplência dos clientes, no sentido que saber a idade de um cliente não é um indicativo da propensão desse cliente se tornar inadimplente. Já para a variável atraso máximo, identificamos que 75% dos clientes inadimplentes, pagaram pelo menos o valor mínimo da fatura ou apresentaram atraso no pagamento. De modo geral, para amostra em estudo, notamos ser comum o não pagamento do valor integral da fatura.

A partir da Figura 4.3 notamos um comportamento similar entre os clientes adimplentes e inadimplentes para a variável proporção limite x pagamento, no qual para maioria dos clientes de ambas as classes tal proporção varia próxima de 0 e 0,1, com a presença de uma grande quantidade de *outliers*. Tal informação nos indica que tanto clientes adimplentes quanto inadimplentes realizam pagamentos bem menores que o limite disponibilizado. Entretanto, analisando na ótica de consumo, com a variável proporção do limite utilizado, observamos que os clientes tendem a ter maior variação dessa proporção,

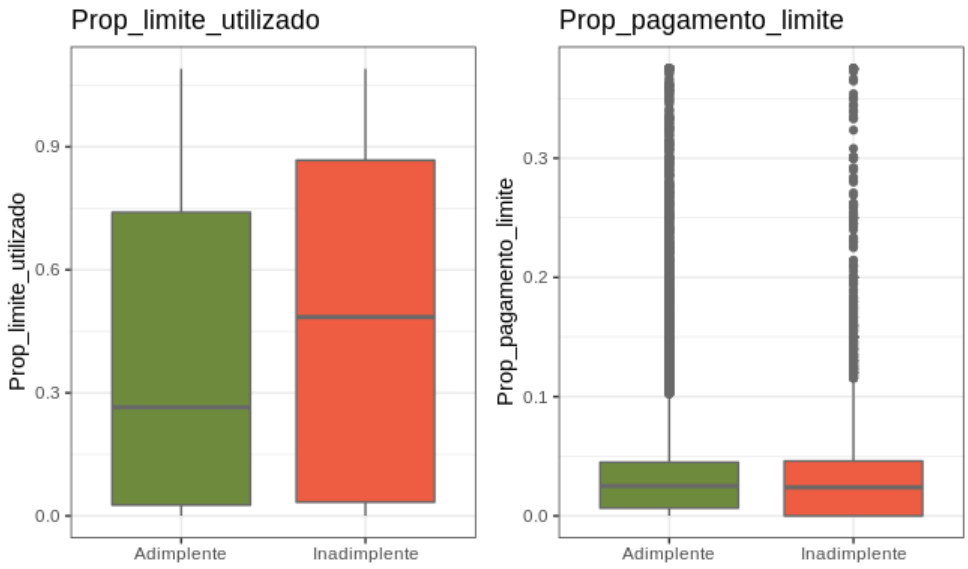

Figura 4.3: Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para a variável Proporção do limite utilizado. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Proporção do pagamento em relação ao limite.

para as duas classes analisadas, embora os clientes inadimplentes tendem a ter um gasto mais próximo do limite disponibilizado, dado que 50% dos clientes inadimplentes utilizam em média metade do limite disponível por fatura. Para a mesma proporção de clientes na classe adimplente esse valor é próximo de $0,25$.

De acordo com a Figura 4.4 destacamos o poder de discriminação da variável quantidade de atraso, de modo que dos clientes inadimplentes, apenas 25% não apresentaram nenhum período de atraso no pagamento da fatura de cartão de crédito, enquanto que para classe adimplente quase todos pagaram suas contas em dia, com exceção de poucos clientes, representados como *outlier* no boxplot, que apresentaram pagamentos com atraso. Em relação a variável proporção de pagamento, identificamos que os clientes adimplentes apresentam uma grande variação da proporção entre o valor pago e o valor da fatura, de modo que 50% dos clientes se concentram - aproximadamente- na faixa de proporção entre $0,20$ e $0,80$, ao passo que para classe oposta (clientes inadimplentes), essa variação é menor. Além disso, a proporção de pagamento para 75% dos clientes indimplentes é de apenas 0,25, ou seja, do montante total que deveria ser pago da fatura - nos meses de referência dessa variável - apenas um quarto desse valor é liquidado pelo cliente.

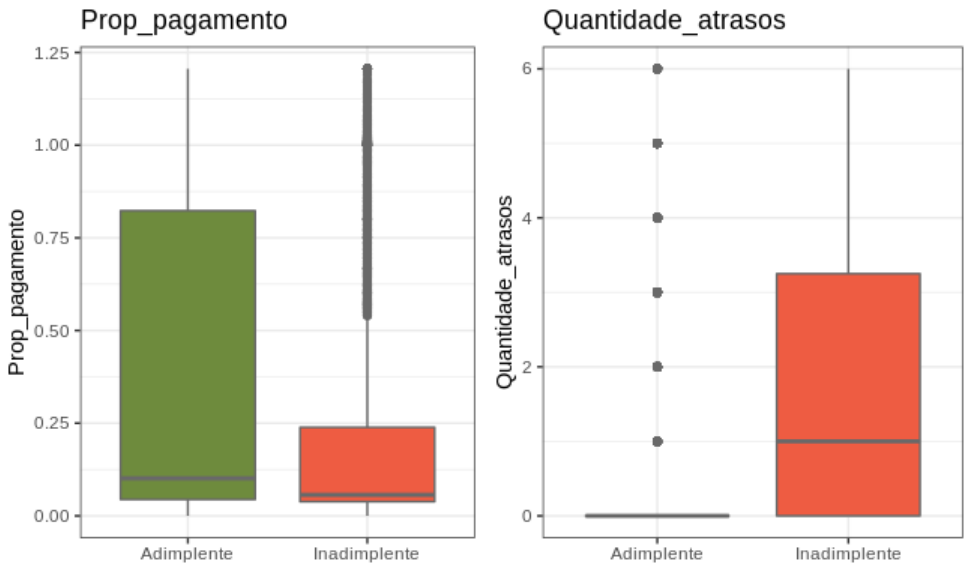

Figura 4.4: Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para a variável Proporção de pagamento. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Quantidade de atrasos.

4.1.2 Resultados

Nesta etapa, realizamos a aplicação das metodologias discutidas anteriormente no conjunto de dados em estudo, Inadimplência de clientes de cartão de crédito. Inicialmente, padronizamos o conjunto de dados, com intuito de deixar todas variáveis, com exceção da variável resposta, em uma mesma escala. Nesse processo, as variáveis são transformadas para terem média zero e desvio padrão um. Em seguida, dividimos o conjunto de dados em dois subconjuntos denominados de conjunto de treinamento e conjunto de teste. A composição desses subconjuntos foi definida a partir de uma seleção aleatória sem reposição do conjunto de dados original, de modo que no conjunto de treinamento foram selecionados 70% das observações pertencentes ao conjunto de dados inicial e no conjunto de teste os 30% restantes. Vale ressaltar que tal seleção foi realizada resguardando as proporções iniciais de clientes adimplentes e inadimplentes da base de dados. A finalidade da divisão do conjunto de dados em dois subconjuntos é para termos insumo para avaliar a performance do classificador em uma base independente, no sentido que no conjunto de treinamento iremos treinar o classificador e no conjunto de teste iremos medir o seu desempenho na classificação de novas observações.

Frisamos que todos procedimentos de análises foram realizados no *software R*. Para aplicação do classificador de máquina de vetores suporte operamos a função *svm* do pacote

e1071, em que utilizamos como parâmetros o *Kernel* Radial/Gaussiano com $\sigma^2 = 0,166$ (valor *default* da função *svm*, que é calculado como $\frac{1}{NV}$, em que NV é número de variáveis do conjunto de treinamento) e o parâmetro de custo igual a um ($C^- = 1$). Os mesmos parâmetros foram utilizados para aplicação do classificador de máquinas de vetores suporte sensível ao custo, entretanto nesse caso o parâmetro de custo para classe adimplente foi mantido como um, enquanto que o custo para classe inadimplente foi $3,1$ ($C^+ = 3,1$). Esse valor foi encontrado a partir de um *grid* de valores que foram testados, no qual escolhemos aquele que maximizou as medidas MCC e G-Média. Como os resultados dos classificadores de regressão logística e floresta aleatória são dados em termos de probabilidades estimadas, empregamos o ponto de corte de 0,5 para definição das classes. Tal ponto foi escolhido, por ser aquele que se assemelharia aos resultados obtidos com o SVM, caso fossem expressos em termos de probabilidade. Por fim, para balanceamento do conjunto de dados aplicamos a técnica de SMOTE no conjunto de treinamento, definindo $N = 2$ e $K = 5$. Ao final do processo de criação das observações sintéticas, o novo conjunto de treinamento ficou com 30.326 observações, das quais 16.382 (54%) são da classe adimplente e 13.944 (46%) da classe inadimplente. Vale ressaltar que não foi realizado nenhuma alteração no conjunto de teste. Os resultados obtidos com a aplicação dos classificadores no conjunto de dados original e balanceado estão expressos na Tabela 4.1.

Tabela 4.1: Resultados da aplicação dos classificadores no conjunto de teste do conjunto de dados Inadimplência de clientes de cartão de crédito. Nas colunas, estão representados as medidas utilizadas para medir o desempenho dos classificadores, em que ACC = Acurácia, SEN = Sensibilidade, VPP = Valor Preditivo Positivo, VPN = Valor Preditivo Negativo, GMedia = G-Média e MCC = Coeficiente de Correlação de Matthews. Nas linhas estão dispostos os classificadores, os quais foram simbolizados pelas siglas, SVM = Máquinas de Vetores Suporte, RL = Regressão Logística, FA = Floresta Aleatória e SVMSC = Máquinas de Vetores Suporte Sensível ao Custo. Os resultados estão apresentados em dois blocos, dados originais, em que o ajuste foi realizado no conjunto de treinamento sem nenhuma alteração e SMOTE, no qual o ajuste foi realizado no conjunto de treinamento balanceado por tal técnica. A marcação em negrito, destaca qual dos três classificadores (SVM, RL e FA) teve melhor desempenho para as medidas avaliadas em cada um dos blocos.

		ACC	**SEN**	**ESP**	**VPP**	**VPN**	**GMedia**	**MCC**
DADOS ORIGINAIS	**SVM**	0.8025	0.2645	0.9556	0.6291	0.8202	0.5028	0.3145
	RL	***0.8031***	0.2661	***0.9560***	***0.6327***	0.8206	0.5043	0.3173
	FA	0.8006	***0.3103***	0.9401	0.5961	***0.8272***	***0.5401***	***0.3256***
	SVMSC	0.7399	0.6308	0.7710	0.4395	0.88	0.6973	0.3583
SMOTE	**SVM**	0.7423	***0.6278***	0.7748	0.4426	***0.8797***	***0.6974***	*0.3601*
	RL	***0.7642***	0.5599	0.8224	***0.4730***	0.8678	0.6785	**0.3609**
	FA	0.7619	0.5236	***0.8297***	0.4668	0.8595	0.6591	0.3395

Conforme ilustrado na Tabela 4.1, para análise dos dados originais, os classificadores de máquinas de vetores suporte e regressão logística apresentam resultados muito próximos para todas métricas aferidas, enquanto que a floresta aleatória apresentou algumas variações, embora estas não sejam tão discordantes dos resultados obtidos nos outros dois classificadores. Analisando individualmente as medidas de performance, destacamos que os classificadores tiveram boa performance na classificação geral, prevendo corretamente 80% das observações do conjunto de teste, entretanto tal resultado pode mascarar o verdadeiro desempenho dos classificadores, dado que apenas 26% dos clientes que são inadimplentes foram classificados como inadimplentes nos classificadores de regressão logística e SVM, e 31% para floresta aleatória. Em contrapartida, aproximadamente 95% dos clientes adimplentes foram classificados corretamente. Esses valores são refletidos no cálculo da G-Média, em que observamos o valor de $0,50$ para SVM e regressão logística, e um valor pouco maior para o classificador de florestas aleatórias $(0,54)$, pelo fato do último ser mais preciso na classificação de clientes inadimplentes. Além disso, podemos observar que dos clientes classificados como adimplentes, aproximadamente 82% foram classificados corretamente, para todos classificadores analisados. De modo geral, podemos inferir que mesmo em um cenário de desbalanceamento moderado, os classificadores apresentaram enviesamento para classe majoritária.

Quando aplicamos a técnica de máquina de vetores suporte sensível ao custo (SVMSC), o enviesamento foi amenizado, ao ponto que a proporção de clientes inadimplentes classificados corretamente aumentou para 0.63, em contrapartida, a proporção de clientes adimplentes classificados corretamente caiu para $0,77$. Nesse sentido, percebemos que o classificador passou a classificar mais clientes como inadimplentes, em vista que melhorou a classificação nessa classe ao custo de perder o desempenho na classe de clientes adimplentes, de qualquer maneira o ganho na classificação de clientes inadimplentes foi maior que a perda na classificação de clientes adimplentes. Tal comportamento também pode ser notado nos valores de VPP e VPN, em que o VPN aumentou e o VPP diminuiu. Em relação as métricas MCC e GMedia, observamos que o valor destas aumentaram devido ao fato da melhor classificação da classe inadimplente que impacta positivamente nos valores de VP e FN.

Aplicando o classificador de máquinas de vetores suporte no conjunto de dados balanceados, a partir do método SMOTE, obtivemos resultados muito semelhantes aos obtidos com o SVMSC. Embora, tenhamos observado uma certa diferença no resultados

das medidas de performance para os demais classificadores, notamos que as conclusões também convergem para aquelas apresentadas para o classificador de máquina de vetores suporte sensível ao custo. Entretanto, vale ressaltar que dos três classificadores, a floresta aleatória foi aquele que apresentou menor desempenho em classificar corretamento os clientes inadimplentes o que contribui para os menores valores de MCC e G-Média. Além disso, notamos que com o conjunto de dados balanceado, os classificadores melhoram sua predição na classe negativa (clientes adimplentes) e pioram a classificação na classe oposta, como podemos ver na medida VPP (valor preditivo positivo) que indica que dos clientes que os classificadores classificaram como inadimplentes, em média apenas 46% são de fato inadimplentes. No entanto, as métricas mais gerais de desempenho, G-Média e MCC, apresentam seus maiores valores, quando o modelo foi treinado na base balanceada ou quando aplicado o algoritmo SVMSC.

4.2 Inadimplência de crédito

Esse segundo conjunto de dados foi obtido de uma competição publicada em https://www.kaggle.com/competitions/GiveMeSomeCredit/overview/description, cujo objetivo é a construção de um algoritmo que prediz a probabilidade de um cliente vir a se tornar inadimplente. Por se tratar de uma competição e os dados serem disponibilizados pelos organizadores, não conseguimos encontrar muitas informações referentes à origem dos dados, porém fica subentendido que estes são referentes a algumas informações cadastrais e de características de pagamento de clientes de um determinado banco. Nesse estudo, a variável de interesse é se o cliente é inadimplente, logo é representada por uma variável dicotômica (variável que assume apenas dois valores 0 ou 1) e para modelá-la temos à disposição as seguintes variáveis:

- RevolvingUtilizationOfUnsecuredLines: Saldo total em cartões de crédito e linhas de crédito pessoais, exceto imóveis e sem dívidas parceladas, como empréstimos de carro, dividido pela soma dos limites de crédito;

- Age: Idade do cliente;

- NumberOfTime30-59DaysPastDueNotWorse: Número de vezes que o cliente ficou de 30-59 dias atrasado, mas não pior nos últimos 2 anos;

- DebtRatio: Pagamentos de dívidas mensais, pensão alimentícia, custos de vida, divididos pela renda bruta mensal;
- MonthlyIncome: Renda mensal;
- NumberOfOpenCreditLinesAndLoans: Número de empréstimos ativos (parcela como empréstimo de carro ou hipoteca) e linhas de crédito (por exemplo, cartões de crédito);
- NumberOfTimes90DaysLate: Número de vezes que o cliente ficou com 90 dias ou mais de atraso;
- NumberRealEstateLoansOrLines: Número de empréstimos hipotecários e imobiliários, incluindo linhas de crédito para aquisição de habitação;
- NumberOfTime60-89DaysPastDueNotWorse: Número de vezes que o cliente ficou 60-89 dias atrasado, mas não pior nos últimos 2 anos;
- NumberOfDependents: Número de dependentes na família excluindo o cliente (cônjuge, filhos etc.);
- SeriousDlqin2yrs: Cliente com 90 dias de atraso ou pior (Sim ou não).

O arquivo original contém uma base de treinamento com 150.000 observações e uma base de teste contendo 101.504 observações. Para esse estudo iremos utilizar apenas a base de treinamento fornecida, pois a base de teste omite a informação referente a inadimplência, logo a base de treinamento fornecida será nosso conjunto de dados. Dos 150.000 clientes listados na base de treinamento, 139.974 (93.32%) clientes são adimplentes e 10.026 (6.68%) inadimplentes.

Seguindo os mesmos critérios utilizados para o conjunto de dados anterior, elegemos apenas algumas variáveis para seguir com as análises. Dentre as variáveis disponíveis, selecionamos e renomeamos as variáveis da seguinte maneira:

- RevolvingUtilizationOfUnsecuredLines - Prop_limite_utilizado;
- Age - Idade;
- NumberOfTime30-59DaysPastDueNotWors - Atraso30_59 dias;
- DebtRatio - Prop_divida_renda;

- NumberOfOpenCreditLinesAndLoan - Qtd_emprestimos_ativos;
- NumberRealEstateLoansOrLine - Qtd_emprestimo_imobiliario.

4.2.1 Análise descritiva e exploratória dos dados

Assim como na Subseção 4.1.1, realizamos uma análise descritiva e exploratória do dados, em que também foi necessário realizar o tratamento de observações *outliers*, para mitigar uma possível influência de tais observações nos resultados de classificação. Nesse caso, substituímos os valores dos pontos discrepantes pelo valor que representa o quantil $98,5\%$.

Figura 4.5: Correlograma entre as variáveis utilizadas no estudo, em que cada quadrado indica o valor da correlação entre as referidas variáveis. A simbologia V1,...,V6, representa de forma simplificada as seis variáveis selecionadas para prosseguimento das análises, de modo que: V1 = Proporção do limite utilizado , V2 = Idade, V3 = Quantidade de atrasos de 30 a 59 dias, V4 = Proporção da renda comprometida com dívida, V5 = Quantidade de empréstimos ativos e V6 = Quantidade de empréstimos imobiliários.

Pelo correlograma ilustrado na Figura 4.5, destacamos uma correlação moderada e positiva entre as variáveis quantidade de empréstimos ativos e quantidade de empréstimo imobiliário, o que de certo modo, é uma relação esperada dado que empréstimos imobiliários tendem a ficar ativos por um longo período de tempo. Na mesma magnitude, observamos que indivíduos que possuem maiores quantidades de empréstimos ativos,

inclinam-se a ter maior comprometimento da renda com dívida. As demais combinações apresentaram baixos valores de correlação, indicando uma relação de quase independência entre as variáveis.

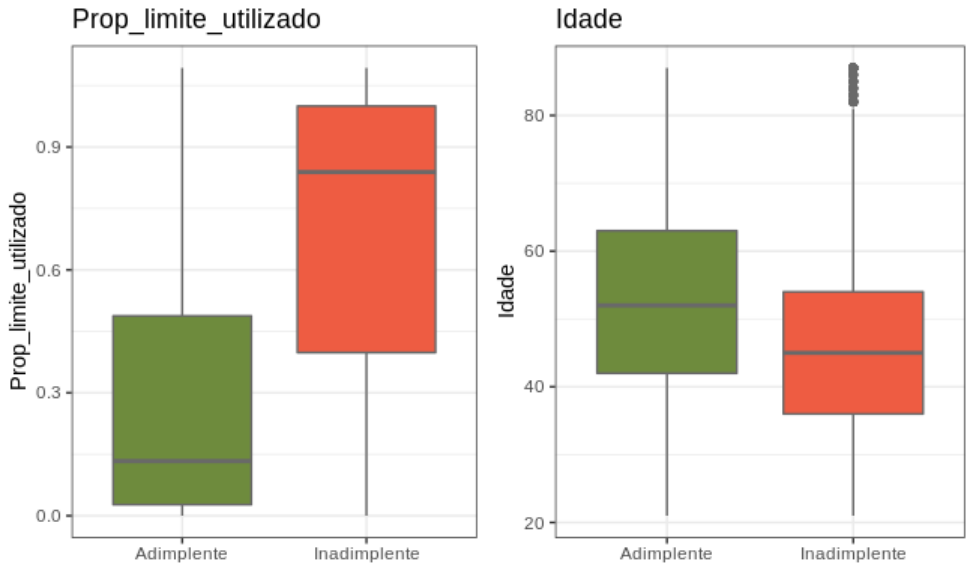

Figura 4.6: Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para a variável Proporção do limite utilizado. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Idade.

A partir da Figura 4.6 destacamos a diferença do comportamento entre os clientes adimplentes e inadimplentes em relação a proporção do limite utilizado, de modo que 50% dos clientes inadimplentes utilizam pelo menos 85% do limite disponibilizado, enquanto que 75% dos clientes adimplentes utilizam menos que a metade do limite disponibilizado. Tal comportamento indica que clientes que utilizam quase todo seu limite estão mais propensos a tornarem-se inadimplentes. Diferentemente do conjunto de dados estudado anteriormente, notamos que para essa amostra de clientes, a idade pode ser um fator que discrimina o comportamento de inadimplência, dado que clientes mais velhos tendem a ter um perfil adimplente.

Pelos boxplots ilustrados na Figura 4.7, observamos que 50% dos clientes inadimplentes apresentaram algum atraso na janela de tempo de 30 a 59 dias, enquanto que os clientes adimplentes, exceto alguns *outliers*, não apresentaram nenhum pagamento com atraso de 30 a 59 dias. Em relação a proporção da renda comprometida com dividas de crédito, notamos que ambos as classes de clientes tem um comportamento parecido, embora a variância dessa variável seja um pouco maior para classe inadimplente.

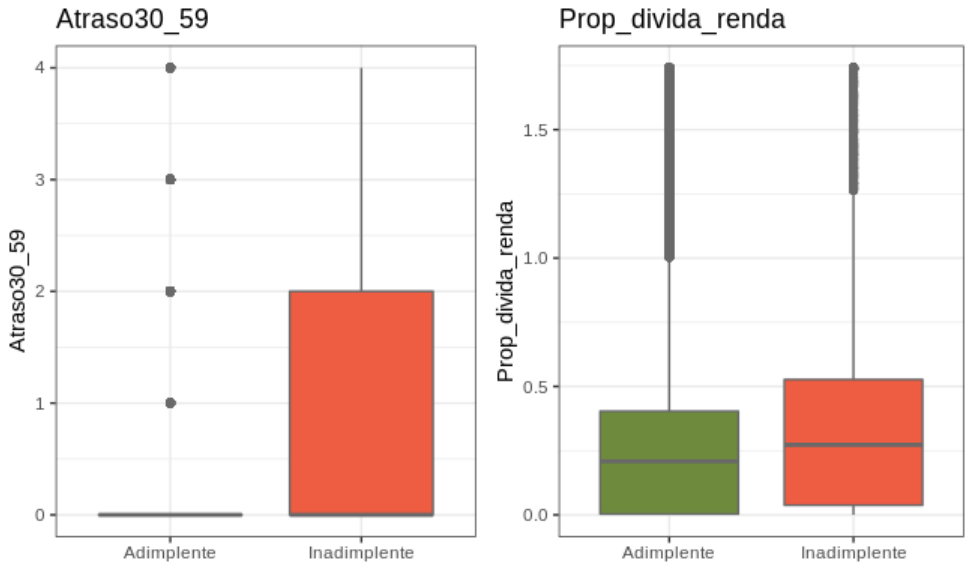

Figura 4.7: Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para a variável Quantidade de atrasos entre 30 e 59 dias. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Proporção da dívida em relação à renda.

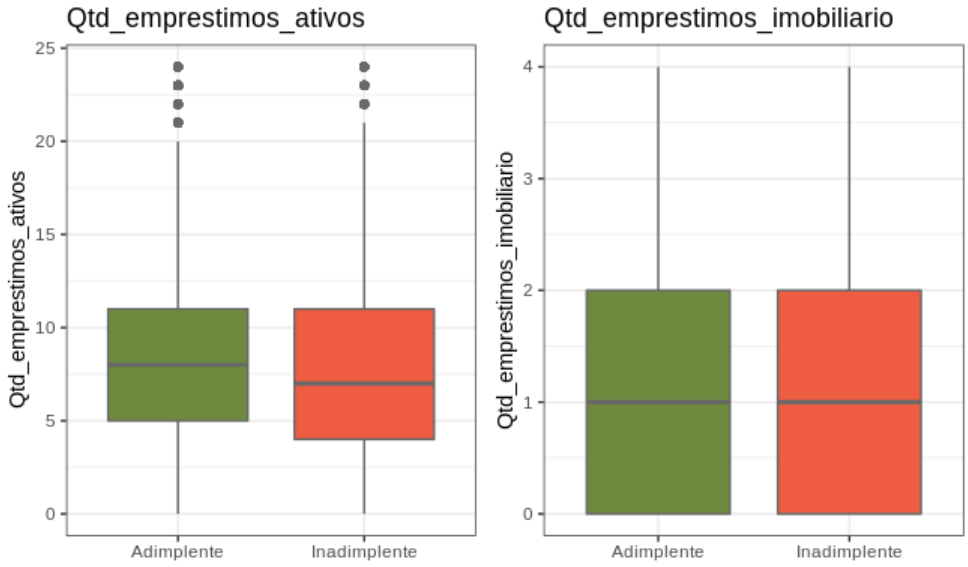

Figura 4.8: Os boxplots do quadro à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) para a variável Quantidade de empréstimos ativos. Enquanto que os boxplots do quadro à direita representam a distribuição dos clientes adimplentes e inadimplentes para a variável Quantidade empréstimos imobiliários.

De acordo com os boxplots apresentados na Figura 4.8, observamos que as variáveis quantidade de empréstimos ativos e quantidade de empréstimos imobiliários, não são variáveis com alto poder de discriminação do comportamento de inadimplência de crédito, dado que comportamento dos clientes adimplentes e inadimplentes são semelhantes para as duas variáveis. Entretanto, ressaltamos que tais variáveis foram mantidas no estudo para fins de investigação.

4.3 Resultados

Previamente à aplicação dos classificadores, abordados nessa monográfica, no conjunto de dados Inadimplência de Crédito, realizamos os mesmos processos descritos na Subseção 4.1.2, no qual apenas alguns parâmetros foram modificados, devido ao desequilíbrio presente nesse conjunto. Para aplicação do classificador de máquinas de vetores suporte sensível ao custo, atribuímos o custo um para classe adimplente ($C^- = 1$) e custo igual a 13 para má classificação da classe inadimplente ($C^+ = 13$). O valor escolhido foi aquele que maximizou as medidas de MCC e G-Média, dentre o *grid* de valores testados. Para balanceamento do conjunto de treinamento via SMOTE, utilizamos $N = 13$ e $K = 5$. Ao final do processo, obtivemos um novo conjunto de treinamento com a volumetria de 195.704 observações, sendo 98.012 da classe adimplente e 97.692 da classe inadimplente, o que equivale a aproximadamente a proporção de 0,5 em cada classe na base de treinamento dos classificadores. Os resultados obtidos nessa etapa para o conjunto de dados Inadimplência de crédito estão expostos na Tabela 4.2.

Analisando os resultados representados na Tabela 4.2 destacamos que todos os classificadores tiverem alto valor de acurácia, o que indica que no geral estão classificando corretamente quase todas observações do conjunto de teste. Entretanto chamamos atenção para o valor dessa medida (ACC), próximo de 93%, que é aproximadamente igual a porcentagem de clientes adimplentes na base treinamento, o que indica o enviesamento do classificador para classe majoritária, classificando quase todas observações nessa classe. Os valores de sensibilidade e especificidade, do classificador de máquinas de vetores suporte, corroboram com tal afirmação, em vista que, apenas 4% dos clientes inadimplentes são classificados corretamente, enquanto que para classe adimplente essa porcentagem é de $99,74\%$. O mesmo comportamento é observado para os classificadores de regressão logística e floresta aleatória, apesar destes terem apresentado desempenho levemente su-

Tabela 4.2: Resultados da aplicação dos classificadores, abordados no estudo, no conjunto de teste do conjunto de dados Inadimplência de crédito. Nas colunas estão representados as medidas utilizadas para medir o desempenho dos classificadores, em que ACC = Acurácia, SEN = Sensibilidade, VPP = Valor Preditivo Positivo, VPN = Valor Preditivo Negativo, GMedia = G-Média e MCC = Coeficiente de Correlação de Matthews. Enquanto que nas linhas estão representados os classificadores, no qual foram simbolizados pelas respectivas siglas, SVM = Máquinas de Vetores Suporte, RL = Regressão Logística, FA = Floresta Aleatória e SVMSC = máquinas de vetores suporte sensível ao custo. Os resultados estão apresentados em dois blocos, dados originais, em que o ajuste foi realizado no conjunto de treinamento sem nenhuma alteração e SMOTE, no qual o ajuste foi realizado no conjunto de treinamento balanceado por tal técnica. A marcação em negrito, destaca qual dos três classificadores (SVM,RL e FA) tive melhor desempenho para as medidas avaliadas em cada um dos blocos.

		ACC	**SEN**	**ESP**	**VPP**	**VPN**	**GMedia**	**MCC**
DADOS ORIGINAIS	**SVM**	0.9328	0.0416	**0.9974**	**0.5474**	0.9347	0.2038	0.1374
	RL	***0.9333***	**0.0987**	*0.9938*	*0.5403*	**0.9382**	**0.3132**	**0.2105**
	FA	0.9319	*0.0846*	0.9934	0.4831	*0.9372*	*0.2899*	*0.1811*
	SVMSC	0.7782	0.7358	0.7812	0.1963	0.9760	0.7582	0.2985
SMOTE	**SVM**	0.7622	***0.7483***	0.7631	0.1866	***0.9766***	***0.7557***	***0.2890***
	RL	*0.7622*	0.7401	0.7637	*0.1854*	0.9758	0.7518	0.2850
	FA	**0.8728**	0.4547	***0.9032***	**0.2544**	0.9579	0.6408	0.2757

perior ao SVM na predição de clientes inadimplentes, em que os valores da sensibilidade foram $9,87\%$ e $8,47\%$, respectivamente. Por conta dessas discrepância na classificação, a medidas MCC e G-Média foram baixas para todos classificadores, apesar de regressão logística ter sido superior aos demais nesses quesitos.

Aplicando a técnica de máquinas de vetores suporte sensível ao custo, conseguimos melhorar a classificação dos clientes inadimplentes, em que passamos de um acerto de aproximadamente 5% para 74%, ao custo da perda, de aproximadamente 23%, da performance na classificação de clientes adimplentes. Tal comportamento é refletido no valor da G-Média que teve uma grande melhora devido a proximidade da proporção de acerto das duas classes em estudo. Apesar disso, notamos que o MCC não teve um aumento expressivo, tal fato está associado a grande quantidade de falsos positivos (FP) que o classificador previu, indicado pelo valor de VPP.

De acordo com a Tabela 4.2, notamos que ao aplicar o método de SMOTE no conjunto de treinamento desbalanceado, os classificadores de máquinas de vetores suporte e regressão logística, aumentaram significativamente o acerto na previsão de cliente inadimplente, obtendo sensibilidade de aproximadamente 74%, enquanto que a especificidade caiu para 76%. Entretanto, para a floresta aleatória esse aumento foi menos expressivo, embora tenha mantida alta taxa de acerto dos clientes adimplentes. Nesse sentido, obser-

vamos que a medida GMédia para floresta aleatória foi menor que os demais classificadores devido a grande diferença entre sensibilidade e especificidade. Em relação as medidas de VPP e VPN, identificamos que os classificadores passaram a errar mais quando classificam um cliente na classe inadimplente, ocasionado um aumento dos falsos positivo, ou seja, após o processo sobreamostragem com o SMOTE, a região de classificação da classe inadimplente passou a ser mais habitada por clientes adimplentes. Tal fato, explica porque o MCC para os classificadores foram baixos.

4.4 Discussão

A partir da análise dos resultados para o conjunto de dados Inadimplência de clientes de cartão de crédito, observamos que o desempenho do classificador de máquinas de vetores suporte e regressão logística foram muito próximos, enquanto que o classificador floresta aleatória apresentou certa diferença, principalmente, na previsão dos clientes inadimplentes, ao qual teve melhor precisão e que resultaram em valores maiores para as medidas G-Média e MCC, quando os classificadores foram treinados na base original desbalanceada. Ao realizar a sobreamostragem no conjunto de treinamento, via SMOTE, notamos que todos os classificadores melhoram seu poder de previsão na classe positiva (clientes inadimplentes), entretanto perderam performance na classificação dos clientes adimplentes. Nesse cenário, o SVM e a regressão logística tiveram performance parecidas, na ótica do MCC, embora o SVM tenha sido superior na métrica G-Média. O classificador de floresta aleatória, diferente do caso anterior (aplicação nos dados originais) em que superou a performance dos demais, foi o que teve menor desempenho nos dados balanceados.

Realizando a análise com o conjunto de dados Inadimplência de crédito, no qual o desbalanceamento é mais severo, identificamos que o classificador que teve melhor desempenho - quando aplicado no conjunto de treinamento desbalanceado - para grande parte das métricas avaliadas foi a regressão logística, seguido do classificador de floresta aleatória. Nesse contexto, o classificador de máquinas de vetores suporte foi o que teve menor desempenho, devido a sua baixa eficiência em predizer corretamente a classe de clientes inadimplentes, apesar dos demais classificadores também terem baixo desempenho nesse quesito. Avaliando os classificadores no conjunto de dados balanceados a partir do método SMOTE, notamos que os classificadores de máquinas de vetores suporte e

regressão logística tiveram resultados muitos próximos para todas as medidas avaliadas, sendo superiores, principalmente para G-Medía, ao classificador de floresta aleatória, que teve baixo desempenho na classificação de clientes inadimplentes quando comparado com aos demais classificadores.

Em suma, observamos que de modo geral o classificador de máquina de vetores suporte teve desempenho semelhante a regressão logística, principalmente em bases de dados balanceadas ou utilizando o SVM sensível ao custo. Em alguns casos teve desempenho levemente superior e em outros inferior. Entretanto, ressaltamos que a utilização da regressão logística tem algumas vantagens em relação ao SVM, como o fato de sua aplicação com variáveis categóricas já estar difundida, além de possibilitar a análise da contribuição de cada variável no resultado obtido, fato esse não disponível no uso do SVM devido a não reversibilidade do dimensionamento do espaço de características. Além disso, a aplicação da regressão logística é menos custosa computacionalmente e temos a possibilidade de analisarmos seus resultados em termos da probabilidade estimada, o que possibilita a escolhe de diferentes pontos de corte para definir as classes.

Em uma breve análise alterando o ponto de corte de classificação da regressão logística e floresta aleatória conseguimos obter resultados um pouco mais satisfatórios que aqueles apresentados neste trabalho. Embora a proposta inicial da metodologia do classificador de máquinas de vetores suporte não nos permitir obter resultados em termos de probabilidades, alguns estudos foram desenvolvidos para esse fim, por exemplo Platt *et al.* (1999) que propôs calcular a probabilidade de um indivíduo pertencer a determinada classe, a partir da aplicação da regressão logística nos resultados obtidos do SVM. Porém, vale a ressalva que tal método pode atribuir mais viés ao resultado final.

Na busca por entender o resultado equiparado entre SVM e regressão logística, e a inferioridade do classificador de floresta aleatória, levantamos a hipótese que tais resultados ocorram devido a quantidade de variáveis explicativas utilizadas para treinar os classificadores, visto que no ajuste da regressão logística o método busca encontrar uma relação linear entre as variáveis, e a partir de tal relação prediz a probabilidade estimada para uma nova observação. Desse forma, acreditamos que a utilização de apenas seis variáveis explicativas favorece a obtenção de uma relação linear entre as variáveis, implicando no bom desempenho da regressão logística. Outro fator que corrobora com tal hipótese, é o bom desempenho do SVM com *kernel* linear, observado em uma breve análise realizada, no qual foi muito semelhante ao SVM com *Kernel* gaussiano para os conjuntos de dados

abordados neste estudo. Nesse sentido, esperamos que um cenário com maior quantidade de covariáveis (como ocorre na classificação de crédito), que dificulte a obtenção de uma relação linear entre elas, os classificadores de máquina de vetores suporte e floresta aleatória tenham melhor desempenho, dado que para o ajuste deste não necessário a linearidade entre as variáveis.

Em relação ao classificador de máquinas de vetores suporte sensível ao custo, observamos que seu desempenho foi muito semelhante ao do SVM com dados balanceados, nos dois conjuntos de dados abordados no estudo. Também ressaltamos, que foi possível observar na prática os pontos teóricos abordados na Subseção 3.1.9, em que aplicando um custo maior para o erro de classificação da classe minoritária, o hiperplano de separação é afastado de tal classe de forma que a região de aceitação para essa classe seja ampliado, contribuindo assim, para a melhora no desempenho de predição dos clientes inadimplentes. Entretanto, tal processo faz com que alguns clientes adimplentes situados próximos a região do hiperplano que antes eram classificados corretamente passam a ser classificados como inadimplentes, conforme pode ser observado na Tabela 4.1 e Tabela 4.2. Dito isso, entendemos que o SVMSC pode ser uma boa metodologia para ser aplicada em dados desbalanceados, quando o objetivo for apenas a predição e não tiver o interesse de realizar alteração no conjunto de treinamento aplicando um pré-processamento.

De modo geral, observamos que os classificadores tendem a ser enviesados para classificação na classe majoritária quando aplicados em conjuntos de dados balanceados, tal fato pode ser evidenciado quando há diferença significante nas medidas de especificidade e sensibilidade, as quais medem respectivamente a proporção de clientes adimplentes classificados corretamente e a proporção de clientes inadimplentes classificados corretamente. Além disso, notamos que o grau do enviesamento se torna maior conforme a severidade do desbalanceamento, dado que a sensibilidade foi próxima de 9% no conjunto de Inadimplência de Crédito, e aproximadamente 26% para o conjunto de dados Inadimplência de Clientes de Cartão de Crédito. Ao balancear o conjunto de dados, sobreamostrando a classe minoritária, identificamos uma melhora na predição dessa classe devido ao aumento da proporção de clientes inadimplentes classificados corretamente conforme pode ser visualizado na Tabela 4.1 e Tabela 4.2. Entretanto, também notamos que o desempenho de classificação da classe majoritária diminui, ao passo que clientes adimplentes que anteriormente eram classificados corretamente, após o balanceamento passam a ser classificados como inadimplentes. Tais afirmações, estão em concordância com o declínio das medidas

de especificidade e VPP, para ambos conjuntos de dados. Em relação ao classificador de máquinas de vetores suporte, além do fato dos conjuntos de dados não serem linearmente separáveis, podemos dizer que a ocorrência da diminuição do desempenho de classificação da classe majoritária, quando balanceamos o conjunto de treinamento, se deve ao fato que com o aumento do número de observações da classe minoritária, o erro acumulativo de cada classes tende a ser próximo, ocasionando a movimentação do hiperplano - que antes era próxima da classes minoritária devido ao erro acumulativo ser menor nessa classe - para uma região intermediária entre as classes. Dessa forma, com a movimentação do hiperplano, observações da classe majoritária que eram classificadas corretamente, porém estavam próximas ao hiperplano, passam a ser classificados na classe minoritária após o balanceamento.

Capítulo 5

Considerações Finais

5.1 Estudos Futuros

As conclusões apontadas nesta monografia foram obtidas a partir de análises de conjunto de dados reais aos quais estão suscetíveis a diversos tipos de ruído e viés que podem afetar os resultados das análises. Nesse sentido, fica como sugestão para próximos trabalhos, realizar um estudo de simulação a fim de verificar se resultados obtidos neste trabalho se mantém em outros cenários. Ademais, para aplicação do classificador de máquinas de vetores suporte, utilizamos apenas uma pequena quantidade de variáveis explicativas - sendo estas apenas de natureza numérica - e parâmetros previamente fixados. Logo, pode ser válido testar o desempenho do SVM com uma grande quantidade de variáveis explicativas, além de incluir variáveis categóricas e variar o tipo de *Kernel* e seus parâmetros.

Conforme comentado anteriormente, os resultados do classificador de máquinas de vetores suporte podem ser obtidos como probabilidades, utilizando algumas metodologias presentes na literatura. Entretanto, tal processo, pode acarretar no enviesamento dos resultados por conta de ser uma estimação sobre os valores resultantes do SVM. Dessa forma, em um próximo trabalho podemos avaliar o impacto de tal viés nos resultados do SVM, utilizando métodos de simulação que contemplem diversos cenários.

Até onde foram nossos conhecimentos e buscas para este estudo, não encontramos formas de analisar a contribuição das variáveis, utilizadas como insumo da classificação no resultado obtido pelo classificador de máquinas de vetores suporte. Logo, ficam como atividades para estudos futuros realizar novas buscas ou desenvolver um novo estudo sobre o esse assunto.

5.2 Considerações Finais

Neste trabalho, abordamos a classificação de crédito, ferramenta muito utilizada por instituições financeiras para tomada de decisão sobre a concessão de seus recursos para solicitantes de crédito. Identificamos que tal decisão está cada vez mais embasada em métodos analíticos obtidos a partir de modelos estatísticos, mais precisamente modelos de aprendizado de máquinas. Além disso, observamos que nesse cenário é comum os conjuntos de dados serem desbalanceados, no qual a quantidade de observações de uma classe se sobrepõem a outra.

Nesse sentido, a metodologia desenvolvida no decorrer dessa monografia, focou em estudar o classificador de máquina de vetores suporte - um algoritmo de classificação que se baseia na construção de um hiperplano de separação no espaço das variáveis - assim como métodos para superar a perca de performance com a aplicação do SVM em dados desbalanceados. Tal estudo foi realizado com o objetivo de avaliar o desempenho do classificador de máquina vetores suporte com conjunto de dados desbalanceados e balanceados, e compará-lo com outros classificadores habitualmente utilizados na classificação de crédito, como regressão logística e floresta aleatória. Para este fim, aplicamos os classificadores em dois conjuntos de dados de crédito, com diferentes proporções de desequilíbrio de classes.

De modo geral, a partir dos resultados obtidos, notamos que o desempenho do classificador de máquinas de vetores suporte ficou pareado com o da regressão logística, em ambos conjunto de dados analisados, e em alguns casos foi superior ao classificador de floresta aleatória. Entretanto, apesar do SVM ter tido performance equiparado com a regressão logística e, em alguns casos, superior a floresta aleatória, sua utilização deve ser avaliada, dado que apresenta algumas desvantagens quando comparado com os demais classificadores, como o custo computacional para convergência do algoritmo, que implica em um maior tempo para obtenção da predição. Além do fato, de não ter uma metodologia difundido para avaliação da importância e interpretação das variáveis utilizadas no estudo. Todavia, o algoritmo do SVM mostrou-se eficiente para lidar com o problema de desequilíbrio de classes quando aplicado com diferentes valores para o parâmetro de custo do termo de penalização, sendo uma boa alternativa ao pré-processamento do conjunto de treinamento.

Por fim, ressaltamos que as conclusões obtidas nesse trabalho estão embasadas nos

conjuntos de dados utilizados para aplicação, podendo ser diferentes caso analisadas em outros cenários. Dessa forma, destacamos que o classificador de máquina de vetores suporte é um método promissor, com forte embasamento teórico, que resulta em um bom desempenho de classificação, além de possibilitar a utilização de um algoritmo capaz de lidar adequadamente com conjunto de dados que apresentam severo desequilíbrio entre as classes. Porém, frisamos que seu uso deve ser avaliado de acordo com o cenário ao qual deseja-se aplicar tal técnica.

Referências Bibliográficas

Akbani, R., Kwek, S. e Japkowicz, N. (2004). Applying support vector machines to imbalanced data sets. volume 3201, páginas 39–50. ISBN 978-3-540-23105-9.

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O. e Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, **94**, 164–184.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, **23**(4), 589–609.

Altman, E. I., Marco, G. e Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the italian experience). *Journal of Banking & Finance*, **18**(3), 505–529.

Arminger, G., Enache, D. e Bonne, T. (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks. *Computational Statistics*, **12**(2).

BACEN, B. C. D. B. (2019). Estatísticas monetárias e de crédito. Acesso em 23-02-2022.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. e Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, **54**(6), 627–635.

Batuwita, R. e Palade, V. (2013). Class imbalance learning methods for support vector machines. *Imbalanced Learning: Foundations, Algorithms, and Applications*, páginas 83–99.

BLATT, A. (1999). *Avaliação de risco e decisão de crédito: um enfoque prático.*. Nobel.

BRIGHAM, E. F., GAPENSKI, L. C. e EHRHARDT, M. C. (2005). Administração financeira–teoria e prática. são paulo, atlas, 2001. *BRIGHAM, Eugene F*.

Chawla, N. V., Bowyer, K. W., Hall, L. O. e Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**, 321–357.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, (3), 326–334.

Desai, V. S., Crook, J. N. e Overstreet Jr, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, **95**(1), 24–37.

Durand, D. (1941). *Risk elements in consumer installment financing*. National Bureau of Economic Research, New York.

Gareth, J., Daniela, W., Trevor, H. e Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Gonçalves, A. R. (2015). Máquina de vetores suporte. *Acesso em 24-01-2022*, **21**.

Hachimi, M., Kaddoum, G., Gagnon, G. e Illy, P. (2020). Multi-stage jamming attacks detection using deep learning combined with kernelized support vector machine in 5g cloud radio access networks.

Hastie, T., Tibshirani, R. e Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York. ISBN 9780387848587.

Investopedia (2021). Credit score. Accesso em 24-01-2022.

Izbicki, R. e dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki.

Kubat, M., Holte, R. e Matwin, S. (1997). Learning when negative examples abound. Em *Machine Learning: ECML-97: 9th European Conference on Machine Learning Prague, Czech Republic, April 23–25, 1997 Proceedings 9*, páginas 146–153. Springer.

Li, J., Wei, H. e Hao, W. (2013). Weight-selected attribute bagging for credit scoring. *Mathematical Problems in Engineering*, **2013**.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, **405**(2), 442–451.

Moro, S., Cortez, P. e Rita, P. (2016). *An automated literature analysis on data mining applications to credit risk assessment*. Springer.

Neto, A. A. e Silva, C. A. T. (1997). *Administração do capital de giro*. Atlas.

Ng, A. (2000). Cs229 lecture notes. *CS229 Lecture notes*, **1**(1). Part V.

Phoungphol, P., Zhang, Y. e Zhao, Y. (2012). Robust multiclass classification for learning from imbalanced biomedical data. *Tsinghua Science and Technology*, **17**, 619–628.

Platt, J. *et al.* (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, **10**(3), 61–74.

Tang, Y., Zhang, Y.-Q., Chawla, N. V. e Krasser, S. (2008). Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **39**(1), 281–288.

Thomas, L. C., Edelman, D. B. e Crook, J. N. (2002). Credit scoring and its applications: Siam monographs on mathematical modeling and computation. *Philadelphia: University City Science Center, SIAM*.

TSURU, S. K. e CENTA, S. A. (2009). Crédito no varejo: para pessoas físicas e jurídicas.

Vapnik, V. N. (1998). *Statistical learning theory*. J. Wiley.

Veropoulos, K., Campbell, I. e Cristianini, N. (1999). Controlling the sensitivity of support vector machines. Em *Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden (IJCAI99)*, páginas 55 – 60. Other: Workshop ML3.

Wang, H., Xu, Q. e Zhou, L. (2015). Large unbalanced credit scoring using lasso-logistic regression ensemble. *PloS one*, **10**(2), e0117844.

WESLEY, D. H. (1993). *Credit risk management: lessons for success*. The Journal of Commercial Lending.

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, **15**(3), 757–770.

Wu, G. e Chang, E. Y. (2003). Class-boundary alignment for imbalanced dataset learning. Em *ICML 2003 Workshop on Learning from Imbalanced Data Sets II, Washington, DC*, páginas 49–56.

Yeh, I.-C. e Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems With Applications*, **36**(2), 2473–2480.

Zhou, L., Lai, K. K. e Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications*, **37**(1), 127–133.