UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E TRANSPORTES

TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO

APLICAÇÃO DE MACHINE LEARNING PARA PREVISÃO DE INADIMPLÊNCIA

VINICIUS SFREDO SOKAL LIMA

Orientador: FLAVIO SANSON FOGLIATTO, *Ph. D.*

PORTO ALEGRE
AGOSTO/2023

Resumo

O presente trabalho aplica algoritmos *de machine learning* para prever a inadimplência de clientes de uma empresa brasileira do setor de varejo e identificar quais são as principais variáveis relacionadas à inadimplência. Foi comparado o desempenho dos algoritmos *K-Nearest Neighbors*, *Random Forest*, *Symbolic Regression* e *Support Vector Machine*, além das técnicas de balanceamento de classes SMOTE e IHT. Além disso, foram utilizadas técnicas de seleção de variáveis e validação cruzada. Todo o trabalho foi desenvolvido utilizando a linguagem de programação Python. A partir da medição e análise de diversas métricas de desempenho, a combinação que gerou as melhores previsões foi o algoritmo *Random Forest* com a técnica de balanceamento de classes SMOTE.

Palavras-chave: inadimplência, aprendizado de máquina, Python, validação cruzada, balanceamento de classes, previsão.

1. Introdução

A oferta de crédito é fundamental para o desenvolvimento e crescimento econômico das nações (GERTLER, 2015). No Brasil, em particular, o crédito - na forma de empréstimos, financiamentos e parcelamentos, entre outros - é a única maneira pela qual muitas famílias são capazes de adquirir seus bens. Por exemplo, em 2021, a população brasileira foi a terceira que mais precisava trabalhar para comprar um celular do modelo iPhone 13 (PICODI, 2021). Seriam necessários 79 dias de trabalho para adquirir o aparelho, enquanto que em países como França, Alemanha e Inglaterra, seriam necessários aproximadamente 10 dias. Nos Estados Unidos da América, menos de 6.

No entanto, apesar de sua importância, o crédito também representa um risco para instituições financeiras e para a estabilidade econômica do país em geral, já que a concessão de crédito de maneira exacerbada a clientes sem condições de pagar pode ter consequências catastróficas. Em caso de inadimplência, não há lucro para as instituições financeiras, nem a possibilidade de os endividados realizarem novos empréstimos. A inadimplência, assim, gera prejuízo para os fornecedores de crédito e diminuição do consumo da população, o que, consequentemente, diminui a produção industrial e, em casos mais graves, pode gerar demissões em massa e aumento do desemprego. Esses, entre outros fatores, foram responsáveis, por exemplo, pela crise imobiliária dos Estados Unidos de 2008 (SORNETTE, 2017).

O presente trabalho tem como objetivo desenvolver uma metodologia utilizando técnicas de *machine learning* (ML ou aprendizado de máquina) para analisar o perfil de clientes de uma empresa varejista e identificar quais variáveis (e.g., renda familiar, sexo e estado civil)

são preditoras de inadimplência. Com isso, será possível analisar o perfil de novos clientes e prever, com certo grau de acurácia, se os mesmos serão capazes de pagar suas dívidas. Os algoritmos usados para resolver esse tipo de problema são denominados algoritmos de classificação, pois cada novo cliente será atribuído a uma classe, e.g., inadimplente ou não. Além disso, serão comparados os resultados obtidos através da utilização de diferentes algoritmos de classificação já conhecidos na literatura, como *Random Forest* (BREIMAN, 2001), *Support Vector Machine* (VAPNIK, 1995) e *K-nearest-neighbor* (FIX, 1952).

A empresa objeto de estudo é uma varejista que comercializa materiais de construção, móveis, eletroeletrônicos e eletrodomésticos, entre outros. Ao todo, são mais de 500 lojas nas regiões Sul, Sudeste e Centro-Oeste do Brasil e mais de 6 milhões de clientes atendidos. Além do varejo, a companhia também funciona como uma instituição financeira, pois oferece opções de cartão de crédito, empréstimo e financiamento. Apesar do bom momento em que se encontra, com inaugurações de novas lojas e aumentos graduais de lucro nos últimos anos, a empresa tem sofrido os efeitos do crescimento nas taxas de inadimplência e de endividamento dos brasileiros. Em 2022, 79,3% dos lares brasileiros possuíam dívidas, enquanto que aproximadamente 30% estavam com parcelas atrasadas em seus pagamentos, i.e., em situação de inadimplência. Esse número equivale a mais de 68 milhões de brasileiros (INFOMONEY, 2022).

O trabalho, por sua vez, trará benefícios para a empresa, que atualmente não implementa nenhuma técnica de *machine learning* nos processos de análise de crédito. Uma metodologia eficaz para previsão de inadimplência permitirá que a empresa analise o perfil de futuros clientes e forneça condições diferenciadas para os mesmos de acordo com sua classificação, de modo que clientes com maiores chances de pagamentos em dia possam receber condições mais arrojadas, como parcelamentos mais prolongados e/ou taxas de juros menores.

2. Referencial Teórico

A nomenclatura *machine learning* surgiu pela primeira vez na tentativa de criar um algoritmo capaz de jogar damas (SAMUEL, 1959). O termo se refere a um processo computacional que utilize dados já existentes para realizar predições sem que seja explicitamente programado para obter um resultado em particular. Ou seja, esses algoritmos são inspirados no pensamento humano, adaptando-se automaticamente (i.e., “aprendendo”) à medida que novos dados são adicionados (MITCHELL, 1997). Nos problemas de classificação, esse processo de adaptação é conhecido como a etapa de treinamento, na qual o algoritmo

recebe dados reais e o resultado esperado para eles. Uma vez treinado, o algoritmo é capaz de receber novos dados e realizar estimativas com base no que "aprendeu" com os dados reais (EL NAQA, 2015). Por exemplo, um algoritmo que diferencia gatos de cachorros precisa primeiramente receber dados de gatos e cachorros, devidamente identificados. Durante a etapa de treinamento, o algoritmo passará a identificar automaticamente características em comum entre os animais (peso, altura, comprimento, tamanho do focinho, cor, etc.) e será capaz de receber dados novos não identificados e classificá-los como gatos ou cachorros, de acordo com essas características. De maneira geral, quanto maior for a quantidade de dados para o treinamento, maior será a acurácia preditiva do modelo (ALPAYDIN, 2014).

Apesar de ser um conceito antigo, as técnicas de *machine learning* demoraram para ganhar popularidade pois, na época em que foram desenvolvidas, computadores eram caros, de difícil acesso e poderiam demorar dias para processar algoritmos simples. Já no século XXI, o processamento computacional tem se tornado cada vez mais acessível. Além disso, houve uma explosão na quantidade de dados coletados e disponíveis sobre praticamente qualquer assunto, o que fez com que empresas, órgãos de governo e acadêmicos buscassem maneiras de analisá-los para obter informações relevantes (JORDAN, 2015). Atualmente, existe uma gama de algoritmos desenvolvidos para solucionar uma grande variedade de problemas (HASTIE, 2011). Essa tecnologia já foi aplicada em campos de conhecimento diversos, como reconhecimento de imagens (FUJIYOSHI, 2019), engenharia aeroespacial (AO, 2010), finanças (GYÖRFI, 2012), ecologia (YANG, 2013) medicina (CLEOPHAS, 2013), combate ao crime (LIN, 2017) e entretenimento (GONG, 2007).

Já em relação à inadimplência, que é uma preocupação global entre organizações governamentais e instituições financeiras (LIN, 2012), diversos modelos matemáticos já foram propostos na literatura para identificar suas causas e auxiliar na tomada de decisões para mitigar esses fatores. Muitos desses estudos abordam o mercado norte-americano, especialmente após a crise imobiliária de 2008 (MIAN, 2009; OJHA, 2021). Porém, outros autores já analisaram o mercado europeu (BARBAGLIA, 2020), empréstimos para compras de imóveis residenciais (KIM, 2021), imóveis comerciais (COWDEN, 2019), empréstimos para famílias de baixa renda brasileiras (VIEIRA, 2019) e empréstimos pessoais através de plataformas P2P (*peer to peer*) online (XU, 2019).

Nas subseções a seguir, serão detalhados conceitos importantes em qualquer problema envolvendo o uso de *machine learning* para análises preditivas, assim como um detalhamento dos algoritmos que serão usados no presente estudo.

2.1 Viés e variância

Um dos conceitos mais básicos no entendimento de modelos preditivos é o equilíbrio entre os erros causados por viés e variância, popularmente conhecido como *Bias-Variance Tradeoff*. Primeiramente, é necessário entender a diferença entre os dois termos, apresentada na Figura 1.

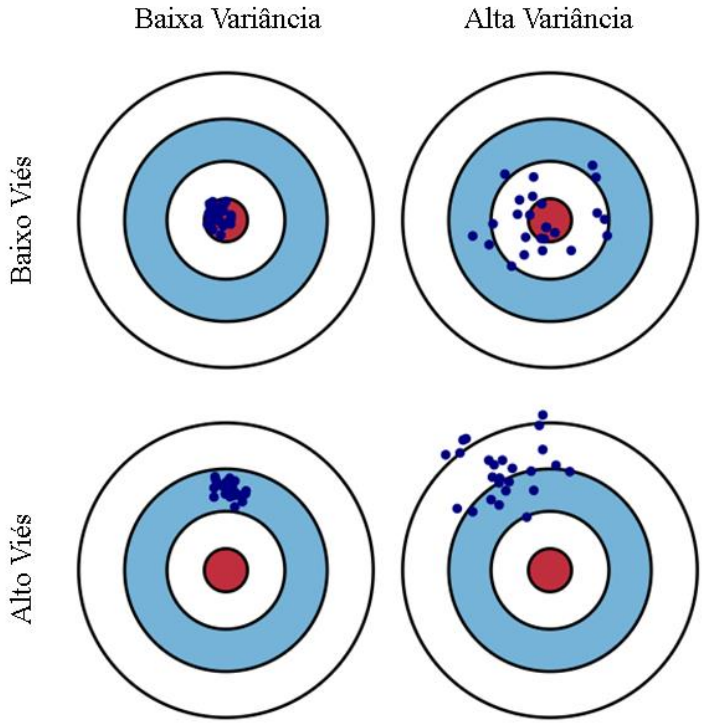

Figura 1: Demonstração dos efeitos de viés e variância. Fonte: adaptado de Fortmann-Roe (2012)

Erros por viés ocorrem quando o modelo está demasiadamente simplificado e não consegue entender as relações entre as variáveis e os resultados esperados. Modelos enviesados tendem a gerar previsões similares, independentemente dos dados analisados. Erros por variância, pelo contrário, acontecem em modelos muito complexos que capturam muitos ruídos presentes nos dados. Nesse caso, a consequência é um modelo que, apesar de adaptado aos dados já existentes, não é capaz de realizar predições assertivas quando é apresentado a novos dados (FORTMANN-ROE, 2012). Em outros palavras, modelos preditivos com alta variância sofrem do fenômeno conhecido como *overfitting* (super ajuste), enquanto modelos com alto viés sofrem de *underfitting* (ajuste insuficiente). Essa diferença pode ser visualizada na Figura 2.

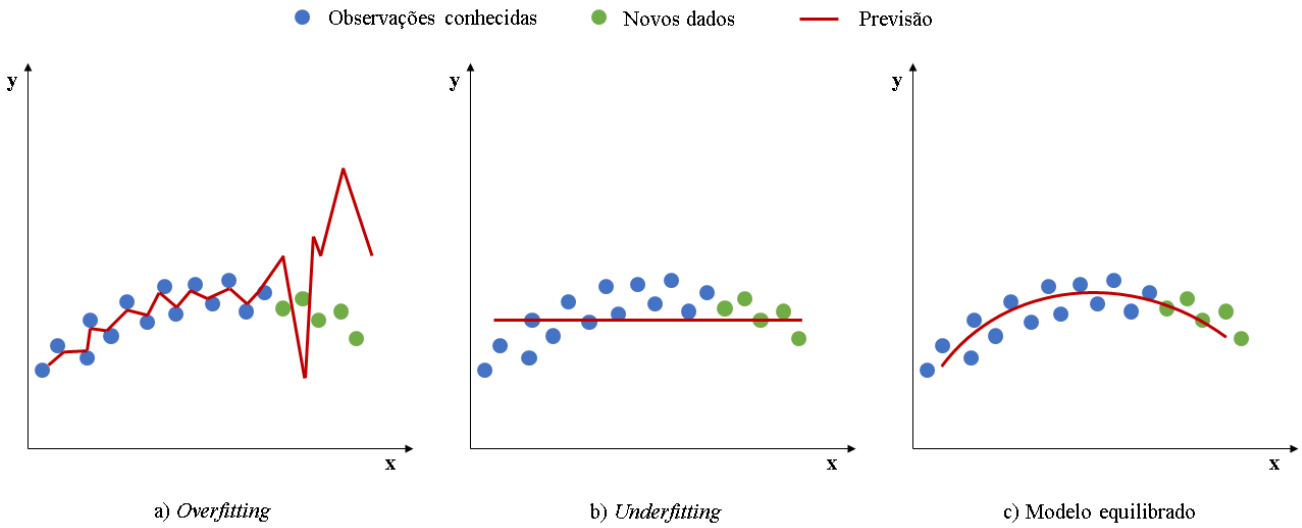

Figura 2: Diferença entre *overfitting* e *underfitting*. Fonte: elaborada pelo autor

É importante que exista um equilíbrio na complexidade do modelo, de modo que ele seja capaz de capturar relações e características importantes nos dados existentes ao mesmo tempo que consiga ser utilizado para prever características de novas observações. Sendo assim, é fundamental que os algoritmos sejam executados mais de uma vez até que se encontre esse ponto de equilíbrio. Adicionalmente, é recomendado o uso de técnicas de balanceamento e validação cruzada (HASTIE, 2009), que serão descritas em seções posteriores.

2.2 Seleção de variáveis

De maneira geral, algoritmos de classificação se beneficiam de uma etapa prévia de seleção de variáveis, já que informações redundantes e/ou sem relação com o problema a ser solucionado podem prejudicar o poder preditivo do modelo (DASH, 2018). Adicionalmente, bancos de dados de menor escala podem reduzir significativamente a capacidade de processamento computacional necessária para executar o algoritmo e, consequentemente, o custo envolvido (NARGESIAN et al., 2017).

Na literatura, as três metodologias utilizadas para seleção de variáveis são conhecidas como *filter* (filtro), *wrapper* (embrulho) e *embedded* (embutimento):

- A abordagem *filter* utiliza métricas estatísticas, como correlação entre as variáveis e o resultado final, para atribuir notas, ou pesos, para as variáveis do problema. Essa alternativa é mais simples que as outras duas e, por isso, demanda menos poder de processamento (SÁNCHEZ-MAROÑO et al., 2007);

- A metodologia *wrapper* é a mais complexa, pois é um processo iterativo que ocorre juntamente com o próprio algoritmo de classificação. A seleção por embrulho costuma obter resultados melhores, porém demanda mais tempo de processamento (EL ABOUDI, 2016);
- Os métodos *embedded* (ou híbridos), por sua vez, são uma mescla dos dois anteriores, resultando em tempos de processamento e acurácia intermediárias (LIU, 2019).

2.3 Balanceamento de classes

Outro ponto que precisa ser considerado em modelos preditivos é o fenômeno conhecido como desbalanceamento entre as classes, i.e., quando uma das classes possui uma proporção muito maior que a outra. Em um banco de dados onde 99% das observações são clientes inadimplentes, por exemplo, qualquer algoritmo que classifique todas as observações como inadimplentes estará certo 99% das vezes, porém não será capaz de classificar corretamente novas observações, já que estará enviesado a classificar todos clientes como inadimplentes. Xue (2015) demonstrou que há uma relação positiva entre o balanceamento das classes de um banco de dados e a capacidade preditiva de um modelo.

Técnicas de balanceamento (ou reamostragem) foram desenvolvidas para harmonizar as proporções entre as classes em estudos preditivos em bancos de dados fortemente desbalanceados. Trata-se de uma etapa que tipicamente ocorre antes da execução dos algoritmos de classificação. As técnicas de balanceamento podem agir aumentando a classe minoritária (*oversampling*), diminuindo a classe majoritária (*undersampling*) ou através de uma combinação de ambas. A alternativa escolhida depende dos dados analisados (SIMSEK et al., 2020). Na sequência, apresenta-se uma técnica de cada tipo (i.e., SMOTE; Chawla et al., 2002, e IHT, Smith et al., 2014), ambas com bons resultados reportados na literatura.

A técnica de *oversampling* SMOTE (*Synthetic Minority Oversampling Technique*) cria observações artificiais para ampliar a classe minoritária. O algoritmo seleciona aleatoriamente uma observação *x* da classe com menos representantes e uma nova observação artificial é criada entre *x* e outras observações na sua vizinhança mais próxima. O processo é repetido até que as classes estejam equilibradas. Alam (2020), Zhu (2019) e Zhou (2012) reportaram resultados superiores utilizando SMOTE comparativamente a outros métodos de reamostragem em problemas relacionados à análise de crédito.

A técnica de *undersampling* IHT (*Instance Hardness Threshold*) atua minimizando a classe majoritária. Esse algoritmo classifica cada observação de acordo com a probabilidade de que ela seja classificada incorretamente, uma propriedade que Smith et al. (2014) denominam de *hardness* (dificuldade). Consequentemente, as observações da classe majoritária com maior valor de *hardness* são eliminadas, pois trariam mais dificuldade para os algoritmos de classificação. Tipicamente, os valores com maior *hardness* (e primeiros a serem eliminados) são *outliers* na amostra. Idealmente, o processo continua até que o número de observações em cada classe seja igual. No entanto, em alguns casos, isso pode resultar na perda de observações importantes da classe majoritária, sendo mais adequado encerrar o algoritmo antes que o equilíbrio total seja atingido (LE et al., 2018). O uso de IHT gerou bons resultados nos trabalhos de previsão de inadimplência reportados por Bastini (2019), Roijmans (2020) e Jin (2021).

2.4 Validação cruzada

Tipicamente, antes de aplicar um algoritmo nas observações sem classificação para prever seus resultados, ele é testado em dados classificados (i.e., com resultados já conhecidos). O banco de dados completo é dividido em amostras menores, conhecidas como porções de Calibração (ou Treino) e de Validação (ou Teste). A porção de Calibração é utilizada para treinar o modelo, que então realiza previsões para os valores da porção de Validação, conforme exemplificado na Figura 3. Como os valores da porção de Validação já são conhecidos, as previsões são comparadas com os valores reais para verificar se o modelo é capaz de gerar boas previsões (MALHOTRA et al., 2020). Nesse simples exemplo da Figura 3, o modelo realizou uma previsão correta e uma incorreta, o que representa uma acurácia de 50%.

Porção	Cliente	Renda	Dependentes	Est. Civil	Inadimplente	
Calibração	A	5000	2	Casado	1	
	B	8000	3	Casado	1	
	C	3000	0	Solteiro	0	Previsão
Validação	D	10000	1	Viúvo	0	1
	E	11000	0	Solteiro	1	1

Figura 3: Exemplo de divisão entre porção de calibração e validação. Fonte: elaborada pelo autor

A técnica mais comum de validação cruzada é a *z-fold*, na qual o banco de dados é dividido em *z* "dobras" (ou porções) menores e *z*-1 dobras são usadas para a calibração do algoritmo, enquanto que a última dobra restante é usada para a validação dos resultados. O

processo é repetido *z* vezes e a acurácia média das iterações é calculada. Essa é uma maneira de minimizar vieses e o impacto de *ouliers* no modelo (RAMEZAN, 2019). O processo completo de avaliação e medição de acurácia dos algoritmos será detalhado na seção de Metodologia.

2.5 Modelos de classificação

Em estudos relacionados a *machine leaning*, diversos autores consideram fundamental que técnicas diferentes sejam comparadas entre si para validar os resultados obtidos (ABU-NIMEH, 2007; BARBAGLIA, 2020; GONG, 2007; NESREEN, 2010; OSISANWO, 2017; VIEIRA, 2019). Neste trabalho, foram selecionadas quatro técnicas para classificar novos clientes da empresa como inadimplentes ou não, as quais são apresentadas a seguir.

2.5.1 *k*-nearest neighbors (KNN)

A técnica *k-nearest neighbors* (KNN), ou *k* Vizinhos Mais Próximos, agrupa as observações de acordo com a proximidade de outras observações com características similares. Quando uma observação nova é inserida no modelo, ela recebe a classificação majoritária de seus *k* "vizinhos" (ou observações) mais próximos. Esse processo pode ser visualizado na Figura 4. Quando *k* = 1, a nova observação será classificada como pertencendo à classe A; no caso de *k* = 3, a classificação seria B.

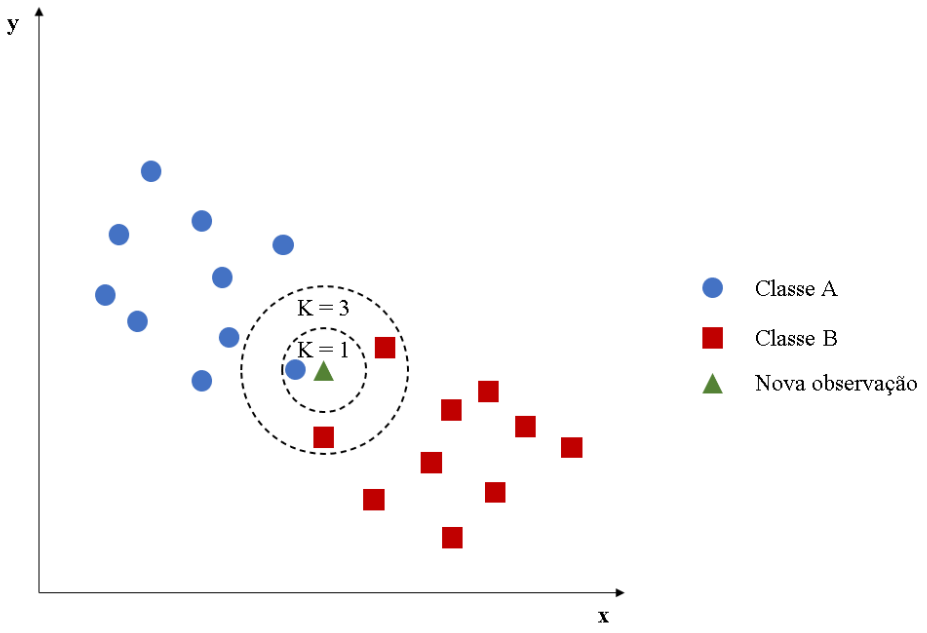

Figura 4: Exemplo de classificação com diferentes valores de *k*. Fonte: elaborada pelo autor

O valor de *k* é selecionado pelo usuário, mas normalmente é determinado através de um processo iterativo de validação cruzada. Valores muito baixos de *k* tendem a capturar muito ruído, enquanto que detalhes importantes e pequenos aglomerados serão perdidos com valores de *k* muito altos (KRAMER, 2013).

2.5.2 Support Vector Machine (SVM)

O algoritmo *Support Vector Machine* (SVM), ou Máquina de Vetor de Suporte, busca encontrar uma linha (ou plano, em caso de problemas em mais de duas dimensões) divisória única que melhor separe as observações do banco de dados em classes distintas, conforme ilustrado na Figura 5. As observações mais próximas da linha divisória são denominadas de vetores de suporte, enquanto a distância entre elas é denominada margem. Infinitas divisórias podem ser traçadas; porém, o algoritmo seleciona aquela que maximiza a margem entre as duas classes (VAPNIK, 1995).

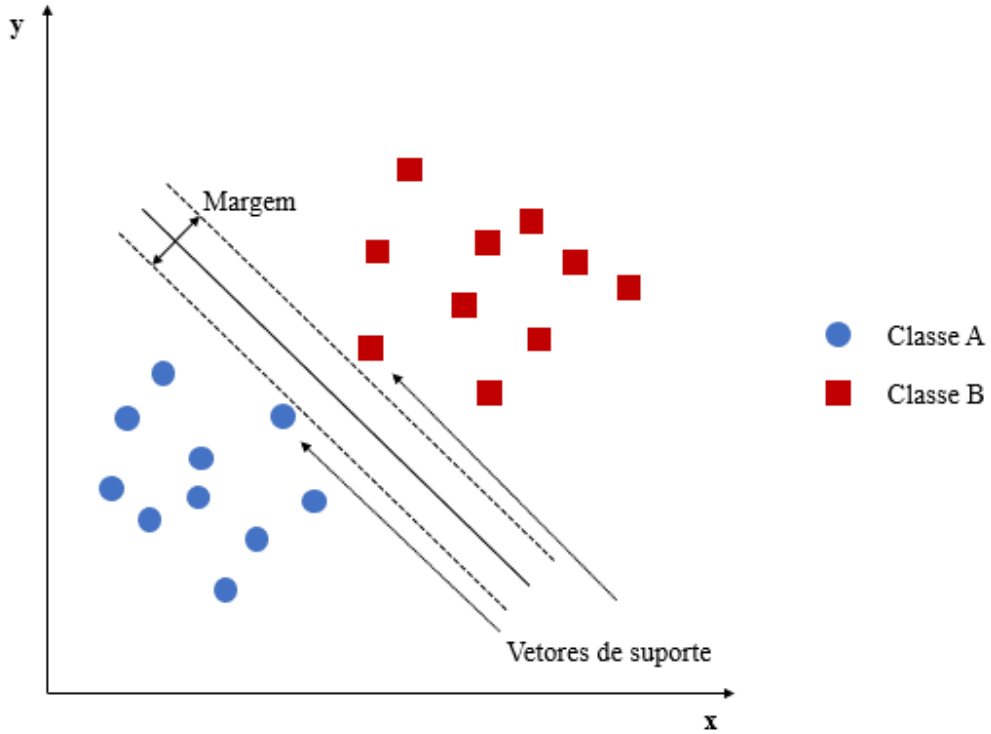

Figura 5: Ilustração do algoritmo SVM. Fonte: elaborada pelo autor

A técnica SVM utiliza um parâmetro *C*, definido pelo usuário, que penaliza observações classificadas incorretamente. Valores baixos de *C* toleram um índice maior de classificações incorretas, levando a problemas de *underfitting*; já valores muito altos de *C* tendem a causar *overfitting*. Assim como o parâmetro *k* da técnica KNN, o valor *C* de SVM ideal é encontrado através de um processo iterativo (LEE, 2017; SUTHAHARAN, 2016).

2.5.3 Symbolic Regression (SR)

De acordo com Koza (1992), o algoritmo SR se diferencia dos modelos de regressão tradicionais por considerar, concomitantemente, relações lineares e não-lineares entre as variáveis avaliadas. A técnica já foi aplicada em diversas situações, e.g., Yamashita et al. (2012) a usaram para prever o público em partidas de futebol, Cai (2006) para modelar uma equação de transferência de calor e Yang (2015) para estimar a taxa de crescimento da produção global de petróleo.

A RS não segue uma estrutura linear pré-definida, diferentemente de outros algoritmos tradicionais. Primeiramente, uma equação (ou grupo de equações) matemática é criada para se adequar ao problema proposto. Através de um processo iterativo, parâmetros da equação são modificados e as soluções são avaliadas de acordo com métricas definidas pelo usuário. O processo se encerra assim que condições especificadas sejam atingidas (Poli et al., 2008).

Não foram encontrados registros na literatura do uso da regressão simbólica para predição de inadimplência de clientes de varejistas.

2.5.4 *Random Forest* (RF)

O algoritmo Random Forest (RF), ou Floresta Aleatória, é um método versátil que pode ser usado tanto em problemas de regressão quanto de classificação. O RF gera aleatoriamente uma grande quantidade de árvores de decisão (Figura 6), onde cada nó das árvores corresponde a realizações das variáveis analisadas no banco de dados.

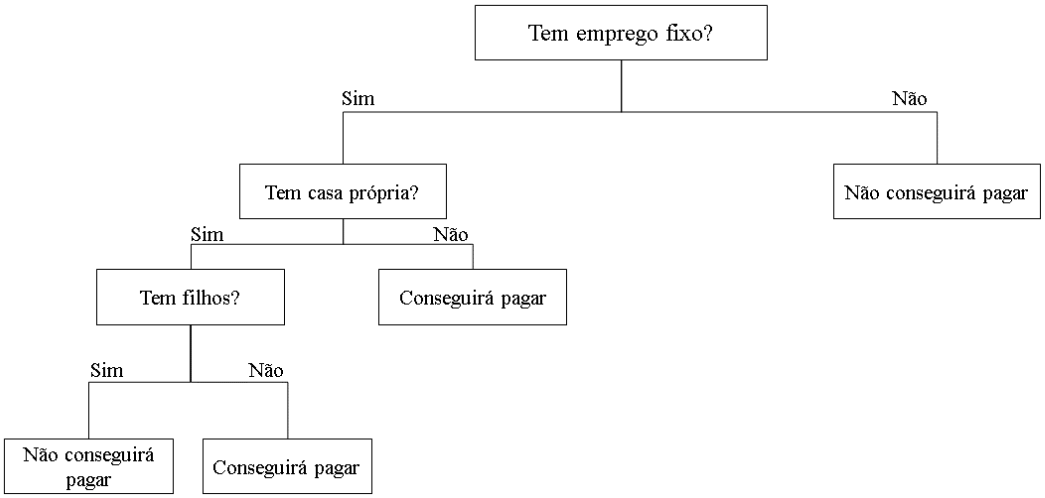

Figura 6: Exemplo de árvore de decisão para um problema de inadimplência. Fonte: elaborada pelo autor

O primeiro nó de uma árvore é denominado de raiz e corresponde a uma das variáveis analisadas. Com base no valor desse nó e uma série de parâmetros, pode ser atribuída uma classificação para a observação ou o nó pode ser subdividido em outros nós (representados por outras variáveis) até que se chegue nos nós finais, também conhecidos como folhas. O parâmetro *d* indica a profundidade máxima, i. e. o número máximo de subdivisões (nós) que o algoritmo pode realizar. Ao final da árvore de decisão, chega-se a uma conclusão a respeito da classificação da observação (no caso da Figura 6, se o cliente conseguirá ou não pagar seu empréstimo). O RF gera um número *n* de árvores de decisão, nas quais a ordem dos nós é escolhida aleatoriamente. Novas observações são testadas em todas essas árvores sendo a classificação final aquela obtida com maior frequência. Zhu (2019) afirma que as vantagens do RF em relação a outros algoritmos são a alta acurácia das previsões, o bom funcionamento (até mesmo em bancos de dados desbalanceados), o baixo poder de processamento necessário e a adaptabilidade do modelo quando há dados incompletos em parte do banco. Uma explicação mais detalhada sobre o algoritmo pode ser encontrada em Breiman (2001).

3. Metodologia

Nesta seção, é realizada uma breve contextualização do cenário da empresa, uma classificação do tipo de pesquisa proposta e a descrição das técnicas, ferramentas e métricas utilizadas para manipular e analisar os dados coletados.

3.1 Cenário

A base de dados utilizada foi fornecida pela empresa estudada que, por questões de sigilo, prefere não ser identificada. Os registros contêm informações coletadas em novembro de 2022 de mais de 6 milhões de clientes. Desconsiderando clientes falecidos, clientes empresariais e clientes com informações incompletas, restam aproximadamente 4,8 milhões, dos quais mais da metade (66,17%) estão com os pagamentos em atraso. Além disso, o valor total das dívidas inadimplentes é de R$700.025.440,41 e a média é de R$216,44. Esses números evidenciam como é importante estudar maneiras de melhorar o processo de análise de crédito dos clientes. Também observa-se que o valor médio é relativamente baixo, o que diferencia este trabalho de demais estudos de análise de crédito, que abrangem dívidas de valores elevados, como financiamentos imobiliários. Além da variável dependente, que indica se o cliente é ou não inadimplente, a base de dados possui 8 variáveis independentes, sendo que 4 são numéricas e 4 são categóricas. Todas as variáveis utilizadas estão apresentadas na Tabela 1.

Tabela 1: Tipo, nome e descrição das variáveis utilizadas

Tipo	Variável	Descrição
Numérica	Dependentes	Número de dependentes
	Parcelas	Número de parcelas da dívida
	Idade	Idade do cliente
	Valor	Valor total da dívida
Categórica	Sexo	Feminino; masculino
	Estado civil	Casado; separado; solteiro; viúvo
	Categoria profissional	Agricultor; aposentado; assalariado; autônomo; liberal
	Categoria da compra	Empréstimo pessoal; financiamento; parcelamento de compra; renegociação de dívida
Dependente	Inadimplência	Inadimplente (1) ou bom pagador (não inadimplente) (0)

3.2 Classificação da pesquisa

O presente trabalho é de natureza aplicada, pois técnicas conhecidas são usadas com o intuito de solucionar um problema real dentro do meio empresarial e financeiro: previsão de inadimplência de clientes (VILAÇA, 2010). Em termos de abordagem, é considerado como quantitativo, visto que não envolve observações e/ou entrevistas e todas as análises realizadas são embasadas em modelos matemáticos e estatísticos (DA SILVA, 2005). Tratando-se de objetivos, é uma pesquisa explicativa, uma vez que o foco principal é entender como características de consumidores influenciam o pagamento de suas dívidas (RAUPP, 2006). Por fim, o procedimento pode ser classificado como um estudo de caso, pois serão analisadas informações já armazenadas no banco de dados da empresa a fim de tirar conclusões relevantes para a mesma (VENTURA, 2007).

3.3 Etapas do trabalho

A seguir, serão descritas as seis etapas realizadas para abordar o problema de previsão de inadimplência de clientes, resumidas na Figura 7.

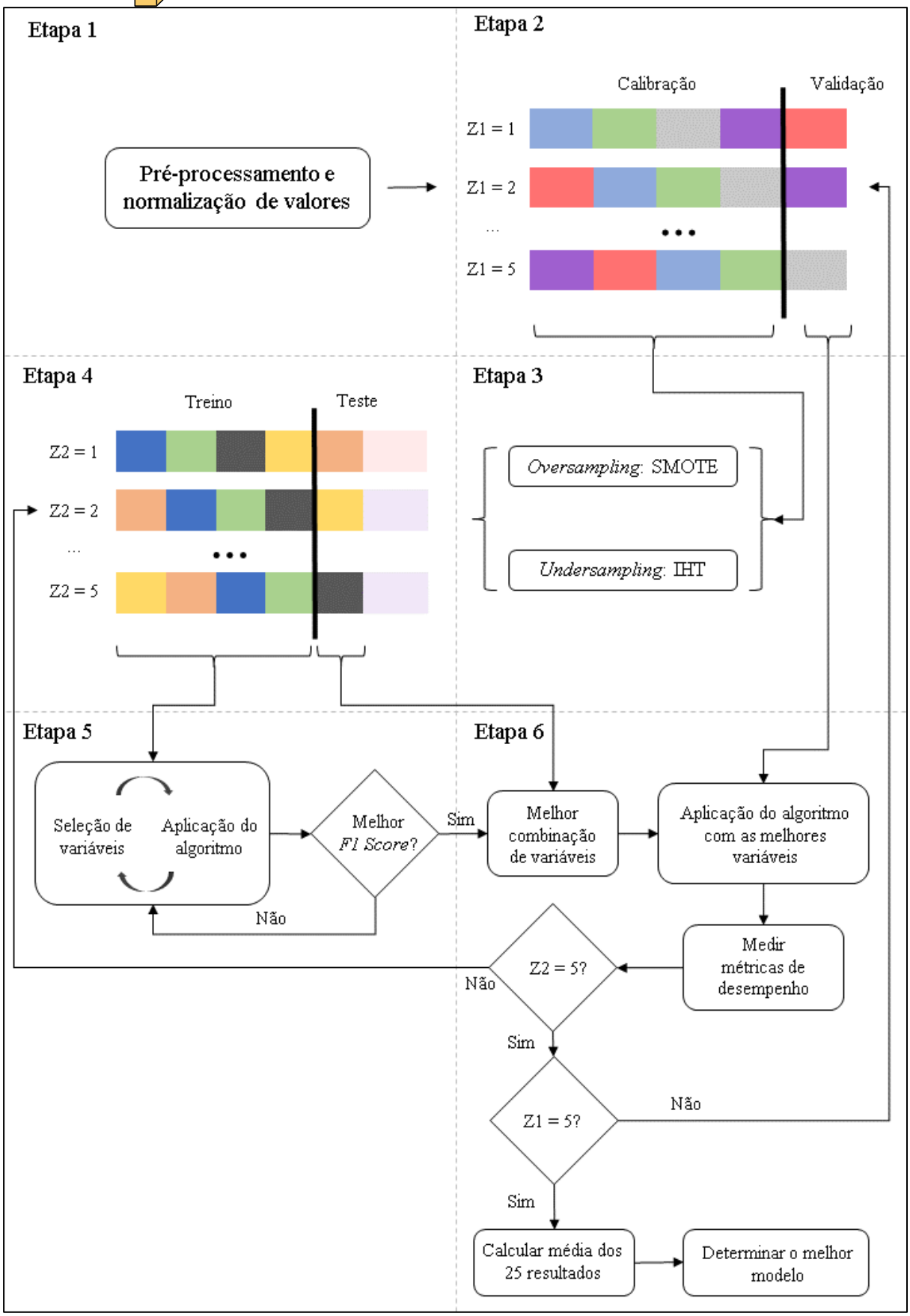

Figura 7: Fluxograma da aplicação do método proposto. Fonte: elaborada pelo autor

3.3.1 Coleta e pré-processamento

Primeiramente, é feita a coleta e pré-processamento dos dados. O pré-processamento começa com a remoção de algumas observações, e.g., clientes já falecidos, clientes empresariais e dados incompletos. Com isso, restam 4.888.057 contas para a análise. Somente foram selecionadas transações nas quais o crédito foi fornecido de alguma maneira, seja em forma de empréstimo ou compra no cartão de crédito. Sendo assim, compras à vista não estão presentes no banco de dados. Além disso, é necessário normalizar as variáveis numéricas para que seus valores fiquem dentro do intervalo [0, 1]. Esse passo é importante para que variáveis de ordens de grandeza diferentes, e.g., renda mensal e idade, sejam consideradas com a mesma relevância pelos algoritmos (AL SHALABI, 2006). A normalização é um cálculo simples apresentado na Equação (1), onde X_{max} e X_{min} são os maiores e menores valores da variável, respectivamente.

$$X_{norm} = \frac{X - Xmin}{Xmax - Xmin} \qquad (1)$$

3.3.2 Validação cruzada

Em segundo lugar, acontece a etapa de validação cruzada descrita na seção 2.4. O banco de dados foi dividido utilizando a técnica *z-fold* com valor de *z* igual a 5, de modo que 4 porções foram usadas para a calibração do algoritmo e 1 foi usada para a validação dos resultados. O valor de *z* foi escolhido por oferecer bom equilíbrio entre os fenômenos de *underfitting* e *overfitting* (BERRAR, 2018; KOHAVI, 1995), apresentados na seção 2.1. Cada uma das porções manteve a proporção da amostra original de contas inadimplentes e bons pagadores, aproximadamente 66% e 34%, respectivamente. Essa proporcionalidade foi alcançada através do uso de uma técnica de amostragem aleatória estratificada (NGUYEN et al., 2021).

3.3.3 Balanceamento de classes

O terceiro passo é a aplicação, somente na porção de calibração, de uma das duas técnicas de balanceamento de classes descritas na seção 2.3: SMOTE (*oversampling*) e IHT (*undersampling*). As duas técnicas foram testadas para que seu desempenho possa ser comparado.

3.3.4 Nova aplicação de validação cruzada

No quarto passo, a porção de calibração, já balanceada, é subdividida em porções de treino e teste. A técnica de validação cruzada *z-fold* é utilizada novamente, desta vez, com o

objetivo para determinar a seleção ótima de variáveis (etapa seguinte do método). Assim como no passo 2, foi escolhido um valor de *z* igual a 5.

3.3.5 Seleção de variáveis

A seleção de variáveis, apresentada na seção 2.2, acontece no quinto passo. Esse processo testa todas as combinações de variáveis e seleciona apenas as que têm maior influência no resultado final. A metodologia *wrapper*, que é a mais complexa, costuma gerar bons resultados (LI, 2017), tendo sido, por isso, escolhida. Foram retidas as variáveis com maior *F1-Score*, o que significa forte relação com o resultado da variável dependente (SONG, 2017).

Os algoritmos de classificação KNN, SVM, SR e RF, apresentados na seção 2.5, foram aplicados no banco de dados após a técnica de balanceamento (IHT ou SMOTE). Todo o processo foi realizado através da linguagem de programação Python. Para as técnicas KNN, SVM e RF, foi utilizada a biblioteca gratuita *scikit-learn* (PEDREGOSA et al., 2011), que contém esses e outros algoritmos de *machine learning*. A técnica SR, por sua vez, foi implementada com a biblioteca *gplearn* (STEPHENS, 2022). Conforme mencionado na seção 2.5, essas funções possuem parâmetros que precisam ser estipulados pelo usuário. Sendo assim, várias possibilidades foram avaliadas. A Tabela 2 apresenta os diferentes parâmetros testados para cada algoritmo.

Tabela 2: parâmetros testados para cada um dos algoritmos

Algoritmo	Parâmetros	Valores testados
K-nearest neighbors (KNN)	*k*	1, 5, 10, 15, 20, 25, 30, 35, 40, 50, 100
Random Forest (RF)	*n*	5, 10, 15, 25, 50, 200, 350, 500
	d	Ilimitado
Support Vector Machine (SVM)	*C*	1, 10, 25, 50, 100, 150, 200
	kernel	Padrão
Symbolic Regression (SR)	Funções matemáticas	Adição, subtração, multiplicação, divisão, raiz quadrada, logaritmo e valor absoluto
	População	500
	Iterações	50
	Probabilidades	*Crossover* = 90%; mutação = 1%

3.3.6 Avaliação de desempenho

No sexto e último passo, as melhores combinações de variáveis selecionadas na etapa 5 foram testadas para cada algoritmo nos dados iniciais não balanceados (etapa 2) para prever os valores da porção de validação. O balanceamento não foi feito para que os resultados se aproximem de um cenário real. Os valores previstos são comparados com os valores reais da porção de validação e as métricas de desempenho são calculadas. Uma vez que todas as combinações de *z-folds* propostas nos passos 2 e 4 tenham sido executadas, os algoritmos são comparados para que seja determinado qual modelo melhor se adaptou ao banco de dados e gerou as previsões mais precisas.

3.4 Métricas de desempenho

Para medir o desempenho dos diferentes algoritmos preditivos, os clientes inadimplentes foram classificados como resultados positivos (1) e os clientes com pagamentos em dia (i.e., bons pagadores) como resultados negativos (0). Sendo assim, os quatro resultados possíveis de classificação são:

i. Positivo Verdadeiro (PV): inadimplente classificado como inadimplente;
ii. Positivo Falso (PF): bom pagador classificado como inadimplente;
iii. Negativo Verdadeiro (NV): bom pagador classificado como bom pagador;
iv. Negativo Falso (NF): inadimplente classificado como bom pagador.

Esses quatro resultados são utilizados para calcular Acurácia, Valor Preditivo Positivo, Valor Preditivo Negativo, Sensibilidade, Especificidade, F1-Score, descritos nas Equações (2), (3), (4), (5), (6) e (7), respectivamente (BOTCHKAREV, 2019; WANG, 2013):

$$Acurácia = \frac{PV+NV}{PV+NV+PF+NF} \tag{2}$$

$$Valor\ Preditivo\ Positivo\ (VPP) = \frac{PV}{PV+PF} \tag{3}$$

$$Valor\ Preditivo\ Negativo\ (VPN) = \frac{NV}{NV+NF} \tag{4}$$

$$Sensibilidade = \frac{PV}{PV+NF} \tag{5}$$

$$Especificidade = \frac{NV}{NV+PF} \tag{6}$$

$$F_1\ score = \frac{2\times VPP\times Sensibilidade}{VPP+Sensibilidade} \tag{7}$$

Adicionalmente, foi medido o parâmetro AUC (*Area Under the Receiver Operating Characteristic Curve*). Essa métrica avalia o quanto o modelo é capaz de distinguir as classes. Um AUC igual a 1 significa uma distinção perfeita, enquanto que um resultado abaixo de 0,5 mostra um desempenho pior do que se as previsões fossem feitas aleatoriamente (LING et al., 2003).

4. Resultados e Discussão

A seguir, são apresentados os resultados obtidos após a execução de todas as etapas descritas na Seção 3.3, assim como uma discussão e comparação com outros estudos relacionados ao mesmo tema.

4.1 Resultados

A Tabela 3 e Tabela 4 apresentam, respectivamente, os resultados das métricas de desempenho nas porções de teste e de validação. Nota-se que os algoritmos possuem parâmetros que devem ser selecionados pelo usuário, como os valores *k* e *C* dos algoritmos KNN e SVM. Os parâmetros testados foram listados na Seção 3.3.5 e só serão apresentados os melhores resultados obtidos entre eles. Além disso, considerando que a validação cruzada foi executada duas vezes, nas etapas 2 e 4 da Metodologia, e que o número de *folds* escolhido foi 5 em ambas etapas, cada combinação de algoritmo e técnica de balanceamento foi executada 25 vezes. Sendo assim, os resultados trazidos são a média e desvio padrão das 25 execuções. Os melhores resultados de cada métrica estão marcados em negrito.

Por se tratar de um problema de previsão de inadimplência, é importante ressaltar que os dois tipos de erro possuem níveis distintos de importância. Um erro do tipo Positivo Falso, i.e., um cliente que irá pagar em dia classificado incorretamente como futuro inadimplente, representa um prejuízo menor para a empresa do que um erro do tipo Negativo Falso, i.e., a previsão incorreta de que um inadimplente irá pagar suas dívidas. No primeiro caso, a empresa recebe um valor que não estava esperando receber, o que não representa risco imediato. No segundo caso, porém, a empresa deixa de receber um pagamento com o qual estava contando. Sendo assim, se muitos erros do tipo Negativo Falso ocorrerem, a fornecedora de crédito sofrerá prejuízos, pois estará emprestando dinheiro para clientes que não têm condições de pagar. Diante desse cenário, as métricas mais representativas sob o ponto de vista da empresa são a Sensibilidade, i.e., a proporção de inadimplentes classificados corretamente e Valor Preditivo

Negativo, i.e., a probabilidade de a previsão de um pagamento em dia estar correta. A métrica AUC também é importante, já que reflete a habilidade do modelo de distinguir as classes.

Nos resultados da porção de teste, apresentados na Tabela 3, é possível perceber que os melhores resultados foram obtidos com a Técnica de Balanceamento IHT, onde todos os algoritmos obtiveram valores acima de 90% para as métricas de Valor Preditivo Negativo, Sensibilidade e AUC. No caso da técnica SMOTE, somente a Regressão Simbólica obteve desempenho no mesmo patamar. Ou seja, nesse caso, a técnica IHT se adaptou melhor ao banco de dados.

Tabela 3: Resultados das métricas de desempenho da porção de teste

Técnica de Balanceamento	**Algoritmo**	**Métricas de desempenho - Média (Desvio)**						
		Acurácia	**VPN**	**Sensibilidade**	**VPP**	**Especificidade**	**AUC**	***F1-score***
IHT	**KNN**	**0.98 (0.008)**	**0.98 (0.015)**	**0.98 (0.015)**	0.98 (0.005)	0.98 (0.005)	**0.98 (0.008)**	**0.98 (0.009)**
	Random Forest	0.975 (0.006)	0.969 (0.01)	0.969 (0.01)	0.981 (0.007)	0.981 (0.007)	0.975 (0.006)	0.975 (0.006)
	Regressão Simbólica	0.957 (0.006)	0.923 (0.01)	0.917 (0.012)	**0.996 (0.002)**	**0.997 (0.002)**	0.957 (0.006)	0.955 (0.007)
	SVM	0.938 (0.021)	0.918 (0.037)	0.912 (0.044)	0.964 (0.024)	0.964 (0.024)	0.938 (0.021)	0.936 (0.023)
SMOTE	**KNN**	0.791 (0.02)	0.763 (0.02)	0.737 (0.026)	0.826 (0.023)	0.845 (0.022)	0.791 (0.02)	0.779 (0.022)
	Random Forest	0.814 (0.01)	0.809 (0.011)	0.806 (0.015)	0.82 (0.014)	0.822 (0.017)	0.814 (0.01)	0.812 (0.01)
	Regressão Simbólica	0.955 (0.008)	0.921 (0.013)	0.915 (0.014)	**0.996 (0.002)**	**0.997 (0.002)**	0.956 (0.007)	0.954 (0.008)
	SVM	0.798 (0.009)	0.763 (0.011)	0.731 (0.019)	0.844 (0.018)	0.864 (0.02)	0.798 (0.009)	0.783 (0.01)

Já nos resultados da porção de validação, trazidos na Tabela 4, SMOTE foi a Técnica de Balanceamento que obteve melhores resultados de maneira geral. Portanto, nota-se que um alto desempenho na porção de teste, que representa a calibração do modelo, não necessariamente significa bons resultados nas previsões da porção de validação. A combinação de IHT com KNN teve um dos piores desempenhos na porção de validação, apesar de ter sido um dos melhores modelos na porção de teste. Esse é um forte indício do fenômeno de *overfitting*, exemplificado na Figura 2-a; ou seja, o modelo está calibrado muito especificamente à porção de treino, mas não é capaz de fazer boas previsões. O fenômeno de *overfitting* também pode ser observado no algoritmo Regressão Simbólica, independentemente da técnica de balanceamento utilizada. A combinação de SMOTE com o algoritmo Random Forest, por sua vez, teve o melhor desempenho de todos, resultando em 88,5% de Sensibilidade e valores aceitáveis de VPN e AUC, 77% e 78%, respectivamente. Essa combinação mostrou-

se mais equilibrada, pois não teve um desempenho tão alto quanto as demais na porção de teste, mas desempenhou bem nas previsões.

Tabela 4: Resultados das métricas de desempenho da porção de validação

Técnica de Balanceamento	Algoritmo	Métricas de desempenho - Média (Desvio)						
		Acurácia	**VPN**	**Sensibilidade**	**VPP**	**Especificidade**	**AUC**	***F1-score***
IHT	**KNN**	0.715 (0.019)	0.627 (0.017)	0.634 (0.026)	0.984 (0.013)	0.982 (0.015)	0.779 (0.016)	0.727 (0.022)
	Random Forest	0.707 (0.017)	0.62 (0.015)	0.622 (0.026)	0.983 (0.016)	0.981 (0.018)	0.773 (0.015)	0.718 (0.021)
	Regressão Simbólica	0.663 (0.012)	0.578 (0.009)	0.539 (0.018)	**0.996 (0.006)**	**0.997 (0.005)**	0.743 (0.009)	0.657 (0.016)
	SVM	0.71 (0.057)	0.64 (0.049)	0.627 (0.087)	0.966 (0.04)	0.962 (0.044)	0.766 (0.048)	0.713 (0.072)
SMOTE	**KNN**	**0.765 (0.028)**	0.715 (0.036)	0.821 (0.035)	0.879 (0.022)	0.801 (0.039)	0.774 (0.027)	0.807 (0.025)
	Random Forest	0.79 (0.029)	**0.771 (0.044)**	**0.885 (0.039)**	0.867 (0.023)	0.76 (0.048)	**0.782 (0.029)**	**0.834 (0.025)**
	Regressão Simbólica	0.663 (0.012)	0.578 (0.009)	0.539 (0.018)	**0.996 (0.006)**	**0.997 (0.005)**	0.743 (0.009)	0.657 (0.016)
	SVM	0.764 (0.024)	0.708 (0.029)	0.805 (0.033)	0.891 (0.026)	0.826 (0.048)	0.779 (0.025)	0.803 (0.022)

A estabilidade das previsões também pode ser observada na Figura 8, que apresenta um *box plot* dos valores de sensibilidade obtidos com cada modelo. A combinação de SMOTE com *Random Forest* é a que entrega resultados mais altos, apesar de apresentar alguns *outliers*, posicionados abaixo do primeiro quartil. As combinações KNN/SMOTE e SVM/SMOTE mostraram desempenho levemente inferior, mas também aceitáveis. Além disso, com exceção ao caso de SR, que obteve valores parecidos com as duas técnicas de balanceamento, é possível observar que a técnica SMOTE teve resultados notavelmente superiores aos de IHT.

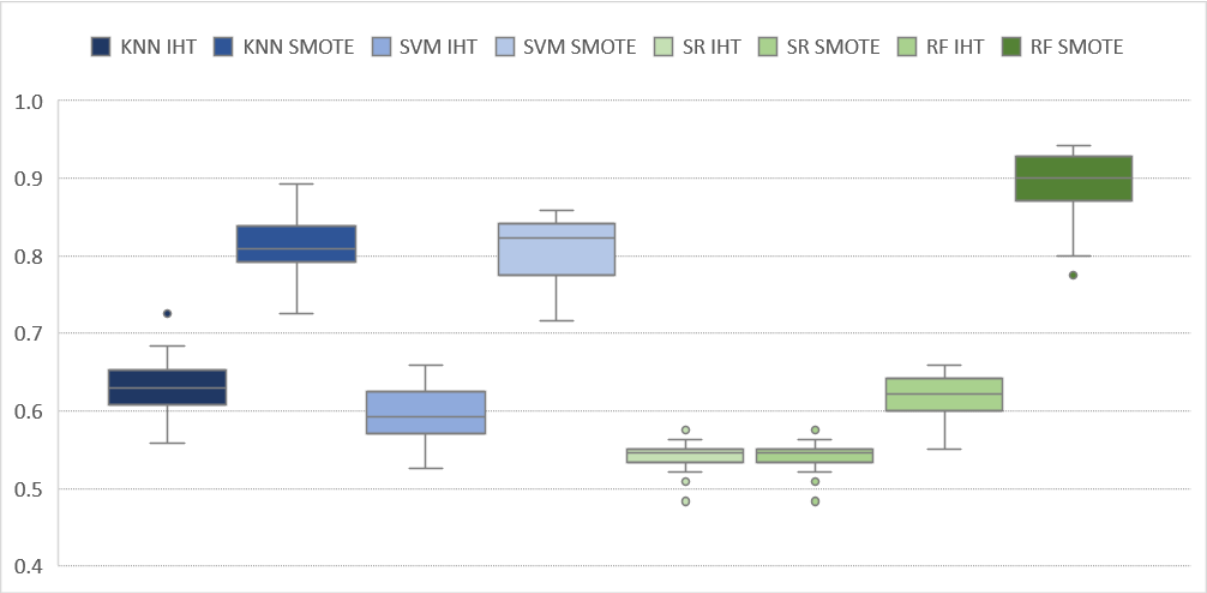

Figura 8: *Box plot* dos valores de sensibildiade da porção de validação

Adicionalmente, a Tabela 5 apresenta quais foram as variáveis selecionadas com mais frequência nas 25 execuções de cada um dos três modelos com melhor desempenho preditivo, KNN/SMOTE, RF/SMOTE e SVM/SMOTE. As variáveis com maior destaque foram ‘renda mensal’ e ‘número de parcelas’, que tiveram uma frequência maior que 23 nos três modelos. Além disso, outras variáveis de alta importância, com frequência acima de 20 em pelo menos 2 modelos, foram: ‘idade’, ‘valor da dívida’, ‘número de dependentes’ e ‘estado civil = separado’. Também observa-se que a combinação com os melhores resultados, RF/SMOTE, foi a que utilizou mais variáveis de maneira geral. A variável menos utilizada nesse modelo foi selecionada 11 vezes, contra 6 e 0 zero vezes nas combinações SVM/SMOTE e KNN/SMOTE, respectivamente.

Tabela 5: Variáveis selecionadas pelos três melhores modelos e suas frequências nas 25 iterações

Variável \ Algoritmo	Frequência de ocorrência		
	KNN/SMOTE	RF/SMOTE	SVM/SMOTE
Idade	5	25	25
nº parcelas	25	25	25
Renda mensal	23	25	25
Valor da dívida	13	25	24
Produto: financiamento	13	24	10
nº de dependentes	0	23	25
Estado civil: separado	0	20	22
Sexo: feminino	9	20	16
Estado civil: casado	1	19	18
Sexo: masculino	6	17	10
Estado civil: solteiro	2	16	11
Profissão: aposentado	1	16	20
Profissão: agricultor	7	14	25
Profissão: autônomo	5	14	24
Produto: compra parcelada	20	13	16
Produto: renegociação de dívida	10	13	24
Profissão: assalariado	1	13	6
Profissão: profissional liberal	13	13	16
Estado civil: viúvo	4	12	25
Produto: empréstimo pessoal	10	11	22

4.2 Discussão

A discussão sobre os resultados passa por três pontos principais: técnicas de balanceamento, validação cruzada e seleção de variáveis.

4.2.1 Balanceamento

O uso de um banco de dados desbalanceado em problemas de previsão pode influenciar os resultados negativamente, visto que os algoritmos tendem a classificar novas observações como pertencentes à classe majoritária. Por esse motivo, técnicas de balanceamento fizeram parte da metodologia proposta nesse estudo. Alam (2020), Zhu (2019) e Zhou (2012) obtiveram resultados superiores com SMOTE em relação a outras técnicas, o que foi ao encontro dos valores obtidos nesse trabalho. Nas Tabelas 3 e 4, é possível observar que a técnica IHT obteve desempenho superior na porção de teste, que representa a calibração do algoritmo, mas não foi capaz de gerar boas previsões, o que pode caracterizar o fenômeno de *overfitting*. SMOTE, por sua vez, teve um desempenho mais equilibrado, i.e., os resultados da porção de calibração não foram tão altos, mas as previsões na porção de validação foram mais assertivas.

4.2.2 Validação cruzada

Uma característica fundamental dos modelos de previsão é que os resultados continuem consistentes quando novas observações forem inseridas. Para isso, é importante que técnicas de validação cruzadas sejam utilizadas para que a divisão do banco de dados nas porções de calibração e validação seja feita com menos vieses nos dados (HEYMAN, 2001). A técnica utilizada mais frequentemente é a *k-folds*, discutida na Seção 2.4. Autores como Berrar (2018) e Kohavi (1995) utilizam 10 *folds* como padrão. No entanto, Sejuti (2023), Marcot (2021), Fushiki (2011) e Madaan (2021) não encontraram ganho significativo de desempenho ao aumentar-se o número de *folds* de 5 para 10, então o valor 5 foi escolhido. Porém, nos casos citados, não foram utilizadas técnicas de balanceamento de classes. Por isso, o presente estudo seguiu a abordagem proposta por Li (2006) e Nasir (2020), na qual o balanceamento é feito juntamente com a validação cruzada para que todos os *folds* tenham a mesma proporção entre as classes. Essa estratégia foi capaz de gerar bons resultados de sensibilidade, VPN e AUC nas combinações SMOTE/RF, SMOTE/KNN e SMOTE/SVM. Nota-se que o melhor resultado de sensibilidade, 88,5%, foi obtido com o algoritmo *Random Forest*, que está em linha com outras comparações de técnicas de *machine learning* para previsão de inadimplência, e.g., Chen (2018), Zhu (2019), Pandimurugan (2022) e Madaan (2021). Além disso, outro ponto positivo do algoritmo é a sua relativa simplicidade e tempo de execução menor do que os outros, algo também constatado por Zhu (2019).

O melhor resultado obtido com a metodologia proposta foi um valor de sensibilidade de 88,5% (SMOTE/RF), que é superior a Chen (2018), 62,4%; Madaan (2021), 74%; e Gautam (2020), 84,5%. Isso demonstra que o modelo é comparável e até mesmo capaz de superar outros estudos de previsão de inadimplência. No entanto, esse resultado de 88,5% foi superado por aplicações como Shingi (2020), 91,4%, e Zhu (2019), 95%, o que demostra que ainda há espaço para melhorias no modelo.

4.2.3 Seleção de variáveis

Variações nas variáveis selecionadas entre os algoritmos são normais, já que estes possuem estratégias de processamento diferentes, o que faz com que seja importante analisar as variáveis selecionadas por mais de um modelo (NASIR et al., 2020). As variáveis apresentadas na Tabela 5 com alta frequência em pelo menos 2 modelos foram consistentes com estudos prévios. ‘Renda’ (GAUTAM, 2020; MA, 2018; SHEIKH, 2020; VIEIRA, 2019; XU, 2019), ‘número de parcelas’ (GAUTAM, 2020; MA, 2018; MADAAN, 2021; ZHU, 2019),

‘idade’ (LAI, 2020; VIEIRA, 2019), ‘valor da dívida’ (BARBAGLIA, 2021; GAUTAM, 2020; MADAAN, 2021; SHEIKH, 2020; ZHU, 2019), ‘número de dependentes’ (DUTTA, 2021; SHEIKH, 2020) e ‘estado civil’ (SHEIKH, 2020; VIEIRA, 2019; XU, 2019) também tiveram forte relação com inadimplência em outros estudos. Além disso, é valido mencionar que existem mais variáveis que podem ter relação com inadimplência, mas que não estavam disponíveis do banco de dados da empresa analisada, como ‘posse de imóvel próprio’ (MADAAN, 2021; ZHU, 2019), ‘taxa de juros’ (BARBAGLIA, 2021; MA, 2018; MADAAN, 2021; VIEIRA, 2019) e ‘grau de escolaridade’ (SHEIKH, 2020; XU, 2019). Além disso, na Figura 9, nota-se que as duas correlações mais fortes com a inadimplência foram com as variáveis ‘número de parcelas’ e ‘renda’, com valores de -0,3 e -0,26, respectivamente. A correlação negativa indica que a inadimplência tende a aumentar conforme essas variáveis diminuem.

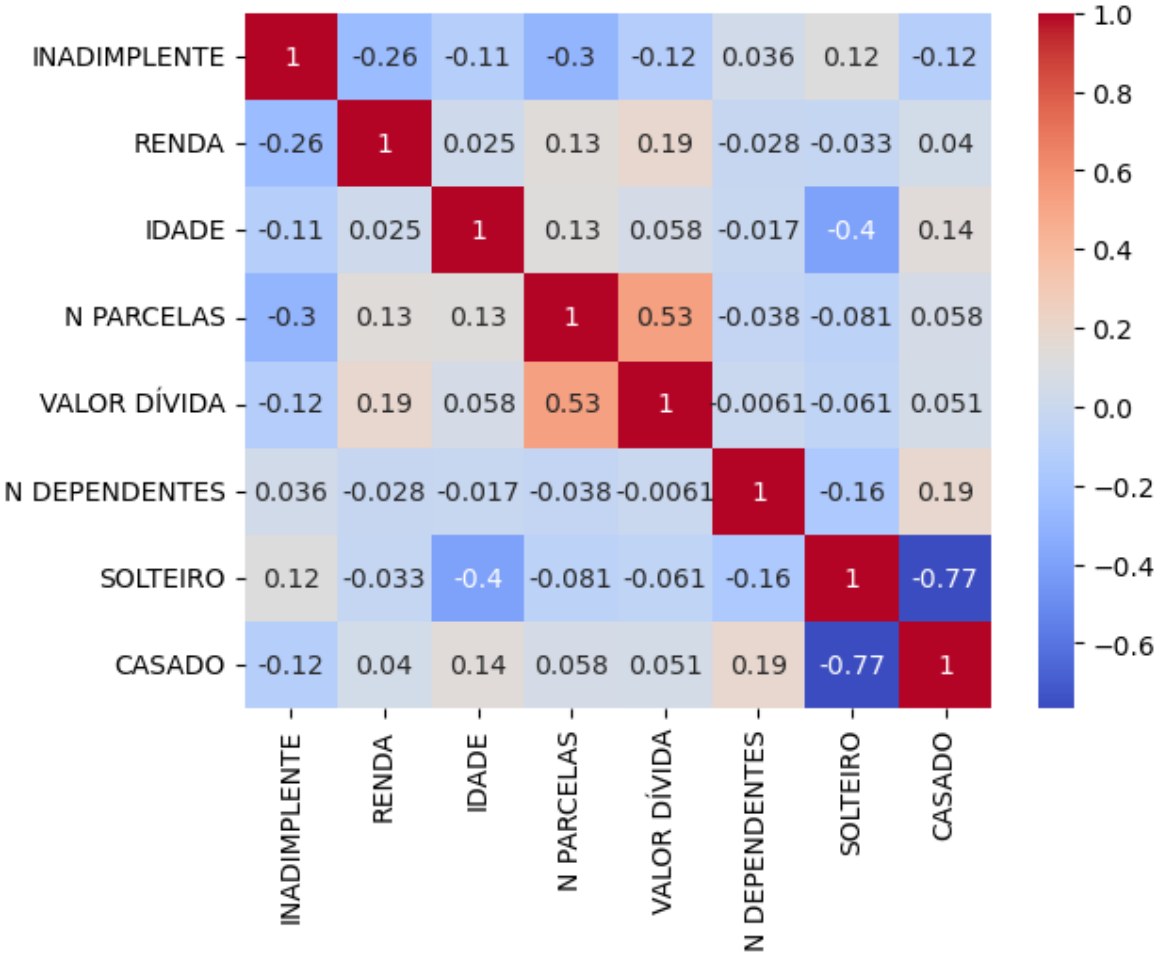

Figura 9: Matriz de correlação entre as variáveis. Fonte: elaborada pelo autor

Percebe-se que apenas algumas informações a respeito da dívida (e.g., valor, parcelas, taxa de juros) e outras a respeito do potencial cliente e sua composição familiar (e.g., renda, número de dependentes e estado civil) são suficientes para gerar boas previsões a respeito da recuperação do capital emprestado. No entanto, por mais que o banco de dados atual tenha sido suficiente para gerar bons resultados, a empresa poderia ampliá-lo e passar a solicitar informações adicionais de seus clientes, como grau de escolaridade e posse de imóvel próprio, mencionados em outros estudos. Tais dados poderiam ser facilmente obtidos durante o cadastro de novos clientes e poderiam ajudar a melhorar o desempenho de futuras previsões.

5. Conclusão

Diante dos resultados apresentados na Seção 4, é possível concluir que o principal objetivo do trabalho foi atingido. A metodologia proposta foi capaz de gerar boas previsões a respeito da inadimplência de clientes com base no banco de dados fornecido pela empresa analisada. Destaca-se que a combinação de técnicas de validação cruzada, balanceamento de classes e seleção de variáveis foram importantes para atingir esse desempenho. As projeções geradas superaram outras aplicações encontradas na literatura, onde essas técnicas não foram implementadas.

Os resultados obtidos têm implicações práticas para a empresa analisada. Algoritmos de *machine learning* são uma maneira rápida, prática e assertiva de gerar previsões a respeito da inadimplência de futuros clientes. Esse conhecimento reduzirá o risco de empréstimos futuros, uma vez que ofertas maiores de crédito podem ser concedidas a clientes classificados como seguros, i.e., aqueles que, segundo projeção do algoritmo, pagarão suas dívidas. Ao mesmo tempo, clientes projetados como inadimplentes podem ter seus empréstimos negados e/ou seus limites de cartão de crédito reduzidos, pois representam um risco financeiro maior. Além disso, esses algoritmos podem ser implementados nos sistemas atuais da empresa para realizar futuras análises de crédito de maneira rápida, automática e imparcial. No entanto, apesar dos benefícios, é importante ter em mente que modelos de *machine learning* não são uma solução estática, i.e., precisam ser constantemente monitorados e ajustados por profissionais qualificados. Afinal, dados financeiros sofrem influência de mudanças macroeconômicas e as variáveis analisadas variam com o tempo (PETROPOULOS et al., 2019).

Estudos futuros poderiam avaliar a possibilidade de classificar os clientes em diferentes categorias de risco de inadimplência, e.g., baixo, médio e alto risco. Isso permitiria à empresa ter uma visão ainda mais detalhada do comportamento de seus clientes e oferecer crédito de

uma maneira mais personalizada para cada um. Além disso, conforme foi visto na Seção 4, ainda há espaço para melhoria nos resultados de acurácia e sensibilidade. O banco de dados fornecido pela empresa possui milhões de clientes; o excesso de variabilidade nos dados dificulta a obtenção de um modelo com boa capacidade preditiva. Um novo estudo poderia ser feito analisando somente uma categoria de dívida, e.g., empréstimo pessoal, compra parcelada ou financiamento. É provável que essas dívidas tenham características diferentes; analisá-las individualmente pode trazer resultados melhores em função da variabilidade menor.

Referências

ABU-NIMEH, S. et al. **A comparison of machine learning techniques for phishing detection**. *In*: **2nd annual eCrime researchers summit**, v. 2, p. 60–69. Nova Iorque: Association for Computing Machinery, 2007.

AL SHALABI, L.; ZYAD, S. Normalization as a preprocessing engine for data mining and the approach of preference matrix. **International conference on dependability of computer systems**, v. 1, p. 207-214., 2006.

ALAM, T. et al. An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. **IEEE Access**, v. 8, p. 201173-201198, 2020.

ALPAYDIN, E. **Introduction to machine learning**, v. 3. Cambridge, EUA: The MIT Press, 2014.

AO, S. I.; RIEGER, B.; AMOUZEGAR, M. **Machine learning and systems engineering**, v. 68. Dordrecht: Springer, 2010.

BARBAGLIA, L.; MANZAN, S.; TOSETTI, E. Forecasting loan default in Europe with machine learning. **Journal of Financial Econometrics**, 2021.

BASTANI, K.; ASGARI, E.; NAMAVARI, H. Wide and deep learning for peer-to-peer lending. **Expert Systems with Applications**, v. 134, p. 209-224, 2019.

BRASIL atinge recordes de 79,3% de famílias endividadas e 30% de inadimplentes. **InfoMoney**, 2022. Disponível em: <https://www.infomoney.com.br/minhas-financas/brasil-atinge-recordes-de-793-de-familias-endividadas-e-30-de-inadimplentes/>. Acesso em 12 de dez. de 2022.

BERRAR, D. Cross-validation. **Encyclopedia of Bioinformatics and Computational Biology**, v. 1, p. 542–545. Amsterdam: Elsevier, 2018.

BOTCHKAREV, A. A new typology design of performance metrics to measure errors in machine learning regression algorithms. **Interdisciplinary Journal of Information, Knowledge, and Management**, v. 14, p. 045-076, 2019.

BREIMAN, L. Random Forests. Machine Learning, v. 45, p. 5–32, 2001.

CAI, W. et al. Heat transfer correlations by symbolic regression, **International Journal of Heat and Mass Transfer**, v. 49, n. 23–24, p. 4352-4359, 2006.

CHAWLA, N. et al. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002.

CHEN, Y.Q., JIANJUN Z.; WING, W.N. Loan default prediction using diversified sensitivity under sampling. **International Conference on Machine Learning and Cybernetics**, v. 1, 2018.

CLEOPHAS, T.J. et al. **Machine learning in medicine**, v. 9. Dordrecht: Springer; 2013.

COWDEN, C. et al. Default Prediction of Commercial Real Estate Properties Using Machine Learning Techniques. **The Journal of Portfolio Management**, v. 45 p. 55-67, 2019.

DA SILVA, D.; SIMON, F. Abordagem quantitativa de análise de dados de pesquisa: construção e validação de escala de atitude. **Cadernos Ceru**, v .16, p. 11-27, 2005.

DASH, M.; LIU, H., Feature selection for classification. **Intelligent Data Analysis**, v. 1, n. 1–4, p. 131-156, 1997.

DUTTA, P. **A** Study On Machine Learning Algorithm For Enhancement Of Loan Prediction. **International Research Journal of Modernization in Engineering Technology and Science**, v. 3, 2021.

EL NAQA, I.et al. **What is machine learning?** *In*: **Machine Learning in Radiation Oncology**, p. 3-11. Cham: Springer, 2015.

EL ABOUDI, N.; BENHLIMA, L. Review on wrapper feature selection approaches. **International Conference on Engineering & MIS**, p. 1-5, 2016.

FIX, E.; HODGES, Jr, J. L. Discriminatory Analysis - Nonparametric Discrimination: Small Sample Performance, **USAF School of Aviation Medicine**, Randolph Air Force Base, Contract No. 41(128)-31, Report No. 11, 1952.

FUJIYOSHI, H.; HIRAKAWA, T.; YAMASHITA, T. Deep learning-based image recognition for autonomous driving, **IATSS Research**, v. 43, n. 4, p. 244-252, 2019.

FUSHIKI, T. Estimation of prediction error by using K-fold cross-validation. **Statistics and Computing**, v. 21, p. 137-146, 2011.

GERTLER, M.; KARADI, P. Monetary policy surprises, credit costs, and economic activity, **American Economic Journal of Macroeconomics**, v. 7, p. 44–76, 2015.

GONG, Y.; XU, W. **Machine learning for multimedia content analysis**. Londres: Springer; 2007.

GYÖRFI, L.; OTTUCSÁK, G.; WALK, H. **Machine learning for financial engineering**, v. 8. Londres: World Scientific, 2012.

HASTIE, T. et al. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, v. 2. Nova Iorque: Springer, 2009.

HEYMAN, R. E.; SLEP, A. M. S. The hazards of predicting divorce without cross-validation. **Journal of Marriage and Family**, v. 63, n. 2, p. 473-479, 2001.

ÍNDICE iPhone 2021: Quantos dias precisamos trabalhar para comprar o novo gadget? **Picodi**, 2021. Disponível em: <https://www.picodi.com/br/mao-de-vaca/iphone-index-2021>. Acesso em 12 de dez. de 2022.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015.

JIN, Y. et al. A Novel Multi-Stage Ensemble Model with a Hybrid Genetic Algorithm for Credit Scoring on Imbalanced Data. **IEEE Access**, vol. 9, p. 143593-143607, 2021.

KIM, D.; SHIN, S. The economic explainability of machine learning and standard econometric models-an application to the U.S. mortgage default risk. **International Journal of Strategic Property Management**, v. 25, n. 5, p. 396–412, 2021.

KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. *In*: **Proceedings of the 14th International Joint Conference on Artificial Intelligence**, v. 2, p. 1137–1143, 1995.

KOZA, J. R. Genetic Programming. **The Programming of Computers by Means of Natural Selection**, v. 33, p. 69–73, 1992.

KRAMER, O. (2013). **K-Nearest Neighbors**. *In*: **Intelligent Systems Reference Library**, v. 51, p. 13-23. Berlin: Springer, 2013.

LEE, Y.C. Application of support vector machines to corporate credit rating prediction. **Expert Systems with Applications**, v. 33, p. 67-74, 2007.

LI, J. et. al. Feature selection: A data perspective. **ACM computing surveys (CSUR)**, v. 50 n. 6, p. 1-45, 2017.

LI, S.T.; SHIUE, W.; HUANG, M.H. The evaluation of consumer loans using support vector machines. **Expert Systems with Applications**, v. 30, n. 4, p. 772-782, 2006.

LIN, Y. L.; CHEN, T. Y.; YU, L. C. Using Machine Learning to Assist Crime Prevention. **6th International Congress on Advanced Applied Informatics**, p. 1029-1030, 2017.

LIN, W. Y.; HU, Y.H.; TSAI, C. F. Machine Learning in Financial Crisis Prediction: A Survey. **Transactions on Systems, Man, and Cybernetics**, v. 42, n. 4, p. 421-436, 2012.

LING, C.X.; HUANG, J.; ZHANG, H. AUC: a statistically consistent and more discriminating measure than accuracy. **International Joint Conference on Artificial Intelligence**, v. 3, p. 519-524, 2003.

LIU, H.; ZHOU, M.; LIU, Q. An embedded feature selection method for imbalanced data classification. **Journal of Automatica Sinica**, v. 6, n. 3, p. 703-715, 2019.

MA, X. et al. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. **Electronic Commerce Research and Applications**, v. 31, p.24-39, 2018.

MADAAN, M. et al. 2021. Loan default prediction using decision trees and random forest: A comparative study. **Materials Science and Engineering**, v. 1022, n. 1, p.12042, 2021.

MALHOTRA, D.K.; MALHOTRA, K.; MALHOTRA, R. Evaluating Consumer Loans Using Machine Learning Techniques. **Applications of Management Science**, v. 20, p. 59-69, 2020.

MARCOT, B. G.; HANEA, A. M. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? **Computational Statistics**, v. 36, n. 3, p. 2009-2031, 2021.

MARQUÉS, A. I.; GARCÍA, V.; SÁNCHEZ, J. S. On the suitability of resampling techniques for the class imbalance problem in credit scoring. **Journal of the Operational Research Society**, v. 64, n. 7, p. 1060-1070, 2013.

MIAN, A.; SUFI, A. The Consequences of Mortgage Credit Expansion: Evidence from the US Mortgage Default Crisis. **Quarterly Journal of Economics**, v. 124, p. 1449–1496, 2009.

MITCHELL, T.M. **Machine learning**, v. 1, n. 9. Nova Iorque: McGraw-Hill; 1997.

NARGESIAN, F. et al. Learning Feature Engineering for Classification. **Twenty-Sixth International Joint Conference on Artificial Intelligence**, p. 2529-2535, 2017.

NASIR, M. et al. A service analytic approach to studying patient no-shows. **Service Business**, v. 14, n. 2, p. 287–313, 2020.

NESREEN K. et al. An Empirical Comparison of Machine Learning Models for Time Series Forecasting, **Econometric Reviews**, v. 29, p. 594-621, 2010.

NGUYEN, T. D. et al. Stratified random sampling from streaming and stored data. **Distributed and Parallel Databases**, v. 39, p. 665-710, 2021.

OJHA, V.; LEE, J. Default analysis in mortgage risk with conventional and deep machine learning focusing on 2008–2009. **Digital Finance**, v. 3, p. 249–271, 2021.

OSISANWO, F. Y., et al. Supervised machine learning algorithms: classification and comparison. **International Journal of Computer Trends and Technology**, v. 48 p. 128-138, 2017.

PANDIMURUGAN, V. et al. Random Forest tree classification algorithm for predicating loan. **Materials Today**, v. 57, p. 2216-2222, 2022.

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

PETROPOULOS, A. et al. A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. Bank for International Settlements, v. 49, 2019.

RAMEZAN, C. et al. Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. **Remote Sens**, v. 11, p. 185, 2019.

RAUPP, F. M.; BEUREN, I. M. Metodologia da pesquisa aplicável às ciências. **Como elaborar trabalhos monográficos em contabilidade: teoria e prática**. São Paulo: Atlas, p. 76-97, 2006.

ROIJMANS, S. **Macroeconomic factors in loan default prediction**. Tese – M.Sc. in Data Science & Society – Tilburg University, Tilburg, 2020.

SAMUEL, A.L. Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210-229, 1959.

SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A.; TOMBILLA-SANROMÁN, M. Filter methods for feature selection: a comparative study. **International Conference on Intelligent Data Engineering and Automated Learning**, 2007.

SEJUTI, Z. A.; ISLAM, M. S. A hybrid CNN–KNN approach for identification of COVID-19 with 5-fold cross validation. **Sensors International**, v. 4, 2023.

SHEIKH, M. A. et al. An Approach for Prediction of Loan Approval using Machine Learning Algorithm. **Proceedings of the International Conference on Electronics and Sustainable Communication Systems**, 2020.

SHINGI, G. A federated learning-based approach for loan defaults prediction. **International Conference on Data Mining Workshops**, p. 362-368, 2020.

SIMSEK, S.; TIAHRT, T.; DAG, A. Stratifying no-show patients into multiple risk groups via a holistic data analytics-based framework. **Decision Support Systems**, p. 132, 2020.

SMITH, M. R. et al. An instance level analysis of data complexity. **Machine Learning**, v. 95, n. 2, p. 225–256, 2014.

SONG, Q.; JIANG, H.; LIU, J. Feature selection based on FDA and F-score for multi-class classification. **Expert Systems with Applications**, v. 81, p.22-27, 2017.

SORNETTE, D. **Why Stock Markets Crash: Critical Events in Complex Financial Systems**, Princeton: Princeton University Press, 2017.

STEPHENS, T. **gplearn Documentation Release** 0.4.2, 2022.

SUTHAHARAN, S. **Support Vector Machine**. *In*: **Integrated Series in Information Systems**, v. 36, p. 207–235. Boston: Springer, 2016.

UNDERSTANDING the Bias-Variance Trade-off. **Scott Fortmann-Roe**, 2012. Disponível em: <http://scott.fortmann-roe.com/docs/BiasVariance.html>. Acesso em 20 de jan. de 2023.

VAPNIK, V. **The Nature of Statistical Learning Theory**. Nova Iorque: Springer, 1995.

VENTURA, M. O estudo de caso como modalidade de pesquisa. **Revista SoCERJ**, v. 20, n. 5, p. 383-386, 2007.

VIEIRA, J. et al. Machine learning models for credit analysis improvements: Predicting low-income families' default. **Applied Soft Computing**, v. 83, 2019.

VILAÇA, M. Pesquisa e ensino: considerações e reflexões. **Revista e-scrita: Revista do Curso de Letras da UNIABEU**, v. 1, n. 2, p. 59-74, 2010.

WANG, C. et. al. Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach. **Diabetes Research and Clinical Practice**, v. 100, n. 1, p. 111-118, 2013.

XU, J. et al. Loan default prediction of Chinese P2P market: a machine learning methodology. **Scientific Reports**, v.11, p. 1-19, 2019.

XUE, J. H.; HALL, P. Why Does Rebalancing Class-Unbalanced Data Improve AUC for Linear Discriminant Analysis? **Transactions on Pattern Analysis and Machine Intelligence**, v. 37, p. 1109-1112, 2015.

YAMASHITA, G. H. et al. Customized prediction of attendance to soccer matches based on symbolic regression and genetic programming. **Expert Systems with Applications**, v. 187. 2022.

YANG, G. et al. Modeling oil production based on symbolic regression, **Energy Policy**, v. 82, p. 48-61, 2015.

YANG, Z.R. **Machine learning approaches to bioinformatics**, v. 4. Hackensack: World Scientific, 2010.

ZHOU, L.; WANG, H. Loan default prediction on large imbalanced data using random forests. **Indonesian Journal of Electrical Engineering**, v. 10, n. 6, p. 1519-1525, 2012.

ZHU, L. et al. A study on predicting loan default based on the random forest algorithm. **Procedia Computer Science**, v. 162, p. 503-513, 2019.