UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIENCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE ESTATISTICA E CIENCIAS ATUARIAIS

Thiago de Jesus dos Santos

APLICAÇÕES DE ALGORITMOS DE *MACHINE LEARNING* PARA PREVISÃO DE INADIMPLÊNCIA EM CONCESSÃO DE CRÉDITO

São Cristóvão – SE

2022

Thiago de Jesus dos Santos

Aplicações de algoritmos de *machine learning* para previsão de inadimplência em concessão de crédito

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Ciências Atuariais.

Orientador (a): Prof. Dr. Carlos Raphael Araújo Daniel

Coorientador (a): Prof. Dr. Cleber Martins Xavier

São Cristóvão – SE

2022

Thiago de Jesus dos Santos

Aplicações de algoritmos de *machine learning* para previsão de inadimplência em concessão de crédito

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como um dos pré-requisitos para obtenção do grau de Bacharel em Ciências Atuariais.

Aprovado em ____/____/____, Nota Final_____.

Banca Examinadora:

_____

Prof. Dr. Carlos Raphael Araújo Daniel

Orientador

_____

Prof. Dr.

Luiz Henrique Gama Dore De Araújo

_____

Prof. Dr.

José Rodrigo Santos Silva

AGRADECIMENTOS

À minha família (Jeisiane Jesus, Erick Saulo, José Neto, Zenilde Mota e Edenilton dos Santos) por todo apoio e pela compressão dos momentos que precisei me ausentar de eventos familiares para dedicar-me aos compromissos acadêmicos. Em especial gostaria de demonstrar minha imensa gratidão a Dona Zenilde Mota de Jesus, minha mãe, por todo amor e cuidado que tem para comigo, saiba que grande parte da minha motivação vem da vontade de proporcionar-te dias melhores. Te amo.

A todos os professores do DECAT, que de maneira direta ou indireta, contribuíram de algum modo para o meu crescimento acadêmico e profissional. Nesses quatros anos como discente deste departamento, pude me conectar e construir boas histórias com todos. Muito obrigado! Gostaria de enfatizar a minha gratidão aos Professores Marcelo, Ulisses, Rodrigo, Raphael e Cristiane (por todos os artigos que elaboramos e publicamos juntos), também gostaria de agradecer às professoras Vanessa e Amanda (pelos eventos e cursos que organizamos), à Professora Wilde do departamento de contabilidade por ter sido tão humana e bondosa no tempo em que fui seu monitor. Enfim, infelizmente não posso falar de todos, senão o texto ficará maior que o TCC rsrs.

Ao reverendíssimo Prof. Dr. Carlos Raphael, que além de meu orientador de iniciação científica e do Trabalho de conclusão de curso, tornou-se meu orientador de vida. Agradeço demais ao universo por ter a sua digníssima pessoa em minha vida. Aproveito também para expressar a minha gratidão ao Prof. Dr. Cleber Martins Xavier por ter contribuído significativamente como coorientador para que essa pesquisa fosse concluída.

A todos os amigos e colegas que a UFS me proporcionou, em especial à Karine de Oliveira, por todas as idas ao Resun, todos os dias de ostentação na moça de Acarajé em que gastávamos o equivalente a dez dias de Resun, todos os açaís após queimar os neurônios nas provas de cálculo, enfim, favela venceu. Também gostaria de agradecer a Jesy Karolayne, seu carinho e apoio em todas as instâncias da vida tem contribuído para meu crescimento pessoal e profissional.

Minha gratidão a todos que passaram em minha vida nessa jornada da UFS e que acreditam no meu potencial, sintam-se representados nesta vitória. Amo vocês.

RESUMO

Devido ao grande avanço computacional, o desenvolvimento de modelos na área de crédito com o intuito de classificar o tipo dos clientes, mensurar a probabilidade de inadimplência e outras informações têm sido sofisticado pelas técnicas de *Machine Learning*. Neste contexto, o presente estudo teve como objetivo o desenvolvimento de modelos preditivos utilizando técnicas de *machine learning*, a fim de identificar os clientes que estão mais propensos a não honrar com suas dívidas perante as instituições financeiras dos Estados Unidos vinculadas à entidade *US Small Business Administration* (SBA). Por meio do estudo descritivo, observou-se um desbalanceamento nos dados ocorrido na variável resposta, referente à inadimplência, pois 82% dos clientes, que gozaram do crédito ofertado pelas instituições, honraram com o pagamento do débito, enquanto 18% tornaram-se inadimplentes. Sendo assim, neste estudo foi proposto que os métodos de regressão logística, *Naive Bayes*, *Decision tree* e *Random Forest* (RF) gerassem modelos treinados em três situações: 1) Cenário real (desbalanceado); 2) Cenário *Undersampling* e 3) Cenário *Oversampling*. Os resultados encontrados apontam que a aplicação das técnicas de *Undersampling* e *Oversampling* ocasionou a redução da acurácia e sensibilidade na maior parte dos modelos, porém gerou um aumento considerável da especificidade de todos os ajustes. Ademais, o *Random Forest* obteve as melhores métricas de avaliação entre os demais algoritmos utilizados, independente do cenário de treinamento proposto. Por fim, utilizando como métrica de avaliação a *Area Under the Curve* (AUC) tem-se que o modelo (M12) gerado pelo algoritmo RF utilizando a técnica de *Oversampling* resultou no melhor desempenho no processo de generalização.

Palavras-chave: Aprendizado de máquina**;** Crédito**;** Inadimplência**;** Balanceamento dos dados; *Random Forest*.

ABSTRACT

Due to the great computational advances, the development of models in the credit area in order to classify, measure the probability of implementation and other ways has been improved by machine learning techniques. In this context, the present study aimed to develop predictive models using machine learning techniques, in order to identify customers who are more likely to default on their debts to US financial institutions linked to a US Small Business Administration entity (SBA). Through descriptive analysis, an imbalance was observed in the data distribution of the response variable, referring to the credit to the implementation by the institutions, since 82% of the customers paid on time, while 18% became defaulter. Therefore, in this study it was proposed that Logistic Regression, Naive Bayes, Decision Tree and Random Forest (RF) methods generated models trained in three situations: 1) Real scenario (unbalanced); 2) Undersampling Scenario and 3) Oversampling Scenario. The results found indicate that the application of balancing techniques cause a reduction in accuracy and sensitivity in the part of the models, with an even larger increase in specificity of all adjustments. In addition, Random Forest obtained the best evaluation measurements among all methods used, regardless of the proposed scenario for the training set. Finally, using the "area under the curve" as an evaluation metric (AUC), the model (M12) generated by the RF model on a sampling technique resulted in the best performance in the generalization process.

Keywords: Machine Learning; Credit; Default; Data Balancing; Random Forest.

LISTA DE ILUSTRAÇÃO

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O risco de crédito pode ser definido como a probabilidade de uma entidade devedora não ser capaz de pagar as suas obrigações, evento comumente conhecido no mercado de crédito por *default*, seu gerenciamento é de extrema importância em organizações do setor financeiro, pois tem como finalidade evitar situações de exposição a tomadores de crédito que não tenham condições de arcar com suas dívidas (FORTI, 2018; MONTEIRO, 2019; PEREIRA, 2020). Em decorrência do tempo, até mesmo clientes considerados bons pagadores, com excelentes históricos de pagamentos, podem em algum momento deixar de cumprir com suas obrigações financeiras, portanto a tomada de decisão em gestão de inadimplência é uma atividade complexa e muitos dos fatores considerados nas análises são difíceis de prever. Assim, os modelos com propósito de identificar os indivíduos propensos à inadimplência estão entre as principais ferramentas para mensurar e mitigar este risco (MONTEIRO, 2019; CORDEIRO, 2020).

Para a mensuração do risco dos clientes em diferentes fases de relacionamento com as instituições credoras, surge o ciclo de crédito que, de acordo com Forti (2018), está dividido em cinco fases: 1) Prospecção: Os modelos desenvolvidos têm como objetivo segmentar o perfil dos clientes para oferta dos produtos; 2) *Credit Score:* nessa fase são desenvolvidos modelos para mensurar a probabilidade de inadimplência dos indivíduos antes de conceder o crédito; 3) *Behavior Score*: busca-se mensurar o risco de inadimplência dos clientes após a concessão do crédito, através do histórico de pagamentos e atrasos, e 4) *Collection Score:* um grupo de modelos com o objetivo de classificar os clientes que já se encontram inadimplentes pelo nível de possibilidade de quitação das dívidas. No presente estudo, o problema abordado está inserido na esteira da fase de *Credit Score.*

Devido ao grande avanço computacional, o desenvolvimento de modelos com o intuito de classificar o tipo dos clientes, mensurar a probabilidade de inadimplência e outras informações têm sido sofisticados pelas técnicas de Aprendizado de Máquina (*Machine Learning* - ML) (SANTOS, 2013; FORTI, 2018), que é um subconjunto da inteligência artificial que busca aplicar técnicas estatísticas para ensinar a máquina a reconhecer padrões e fazer previsões a partir dos dados Mediante a esse aprendizado, os sistemas geram modelos que melhor explicam as informações analisadas, permitindo boas avaliações e tomadas de decisão mais precisas (LI et al., 2020). A aplicação destes métodos pode ser justificada pela necessidade de conhecimento sobre os clientes por intermédio de

dados cadastrais, hábito de pagamento, classificação do risco de ser adimplente ou inadimplente, fraudador ou de boa-fé, recuperado ou não recuperado, entre outros (SANTOS, 2013; FORTI, 2018).

A principal diferença entre os métodos estatísticos tradicionais e as técnicas de ML é que no primeiro caso os modelos baseiam-se em pressupostos feitos pelo pesquisador, como uma correlação ou independência das variáveis, normalidade, entre outros. Já em ML é admissível utilizar algoritmos que aprendam e desenvolvam estruturas específicas para os dados sem a necessidade de suposições (OLIVEIRA, 2020). Deste modo, as empresas estão investindo cada vez mais em aplicações de *machine learning* para que possam extrair o máximo de informações, a fim de auxiliar na tomada de decisão e na definição de processos mais rentáveis. A elaboração de modelos mais robustos e precisos para todas as etapas do ciclo de crédito estão entre os assuntos mais discutidos nos últimos anos (FORTI, 2018).

O presente trabalho está organizado em 6 capítulos, o primeiro refere-se a esta introdução, em seguida tem-se a exposição dos objetivos propostos neste estudo (Capítulo 2). No Capítulo 3 busca-se apresentar um panorama sobre o que os estudiosos têm debatido em relação ao tema abordado. Posteriormente, são explanados os aspectos metodológicos e conceitos importantes para o desenvolvimento dos resultados (Capítulo 4). A exposição e discussão dos resultados estão inseridas no Capítulo 5 e, por fim, as conclusões desenvolvidas são apresentadas no Capítulo 6.

2 OBJETIVOS

2.1 Geral

Desenvolver modelos preditivos utilizando técnicas de *machine learning*, a fim de identificar os clientes que estão mais propensos a não honrar com suas dívidas perante a entidade credora.

2.2 Específicos

- Realizar uma análise descritiva dos dados por meio das variáveis selecionadas para o desenvolvimento dos modelos;
- Explorar as técnicas de balanceamento *Undersampling* e *Oversampling*;
- Desenvolver modelos baseados nos métodos de regressão logística, *Naive Bayes*, *Decision tree* e *Random Forest* em três cenários de treinamento: Original (desbalanceado), *Undersampling* e *Oversampling;*
- Avaliar a performance dos modelos através dos indicadores de acurácia, sensibilidade, especificidade e precisão;
- Comparar e selecionar o melhor modelo para o conjunto de dados utilizado, mediante a análise da *Receiver Operating Characteristic* (Curva ROC) e do indicador Area *Under the Curve* (AUC).

3 REVISÃO DE LITERATURA

Dado a importância do gerenciamento de risco de crédito para as entidades financeiras, diversos estudos estão sendo realizados nas últimas décadas com o objetivo de desenvolver modelos que auxiliem na mitigação dos eventos indesejados, seja através da mensuração da probabilidade de inadimplência nos produtos associados ao crédito, identificação de transações ou clientes fraudulentos, classificação da concessão de crédito, entre outros (CORDEIRO, 2020). Paiva (2015) elaborou um modelo utilizando a regressão logística para prever a inadimplência dos clientes de uma instituição financeira, técnica apontada por Cordeiro (2020) e Forti (2020) como a mais tradicional no mercado.

Andrade (2008) utilizou os dados *Australian Credit Approval* para desenvolver e comparar modelos baseados em alguns algoritmos de *machine learning*. Na sua visão, a utilização dos algoritmos *Boosting e Support Vector Machine* não resultaram melhores desempenhos em comparação com a Regressão logística e *Naive Bayes*. Segundo o autor, observando a complexidade e precisão dos modelos, a Regressão logística e *Naive Bayes* poderiam ser considerados como os que tiveram o melhor desempenho.

Santos (2013) buscou desenvolver modelos baseados nos algoritmos *Naive Bayes*, *decision tree* e redes neurais para os conjuntos de dados *Australian Credit Approval* e *German Credit Data*, disponíveis no repositório da UCI. O interesse do estudo foi classificar os clientes em bons ou maus pagadores, e no que tange ao desempenho das predições, o algoritmo *decision tree* apresentou a maior acurácia para os dados *Australian Credit Approval* e o modelo de rede neural foi o que melhor performou nos dados *German Credit Data.* Ademais, o algoritmo *Naive Bayes* obteve a menor acurácia em ambas as bases de dados.

Forti (2020) em seu estudo aplicou os algoritmos *Random Forest*, *Support Vector Machine* (SVM) e *Gradient Boosting* para um banco de dados real de cobrança, visando identificar os clientes com mais propensão a pagar suas dívidas e comparou com a metodologia tradicional da regressão logística. Utilizando a acurácia como a métrica de avaliação, os algoritmos *Support Vector Machine* e *Gradient Boosting* tiveram um desempenho melhor quando comparados com a Regressão logística, já o *Random Forest* não obteve um poder preditivo superior ao da metodologia tradicional.

Li et al. (2020) geraram modelos para classificar os créditos concedidos por uma instituição financeira, e para os dados utilizados no estudo, o algoritmo *random forest* foi o que apresentou as melhores métricas de avaliação. De maneira similar, Lakshmi e Kavilla (2018) também indicaram o algoritmo *random forest* como o mais acurado aos seus dados, sua análise tinha o objetivo de identificar clientes fraudulentos em uma operadora de crédito. Ainda na linha de pesquisa de identificação de fraudes, Sousa (2021) explorou a aplicação do *Random Forest*, *k-nearest neighbors (*KNN*)* e *SVM* associados a cenários de reamostragem para balanceamento dos dados, pois em temas como fraude é comum encontrar-se bases com as classes desbalanceadas. Seus resultados indicaram que o algoritmo *Random Forest* foi o que melhor se adaptou às situações propostas.

Cordeiro (2020) explorou as técnicas de *Machine learning* para prever o *default* de empresas brasileiras através dos demonstrativos contábeis. Visando o desbalanceamento das classes da variável resposta presente nos dados, o autor utilizou a técnica de combinação *Over + undersampling* para desenvolver os modelos, e identificou que o balanceamento da base teve forte impacto no poder preditivo da regressão logística. Além disso, o *Random forest* apresentou o melhor desempenho independentemente do cenário (balanceado ou desbalanceado).

Em concordância com os resultados de Cordeiro (2020), Stelzer (2020) avaliou 23 técnicas de *machine learning* na abordagem de classificação para pontuação de crédito em diferentes conjuntos de dados, aplicando cinco técnicas de balanceamento. Seus achados apontam que os métodos *ensemble*, que combinam modelos individuais para produzir um melhor modelo (ANICETO, 2016; FORTI, 2018), tiveram desempenhos superiores em comparação com os demais modelos individuais, sobretudo o algoritmo *Random forest*, responsável pelo melhor desempenho em todos os cenários.

Nguyen e Huynh (2020) abordaram o tema sobre desbalanceamento dos dados em cenários de crédito e seus impactos para a técnica de *machine learning.* Os autores compararam os métodos *ensemble* e os individuais em dois conjuntos de dados, sendo que um desses *datasets* está sendo utilizado no presente estudo. Suas análises evidenciaram que os métodos de balanceamento proporcionam uma melhoria significativa no desempenho dos classificadores perante a classe com menor participação, e que os métodos *ensembles* geraram desempenhos melhores que os métodos individuais.

4 METODOLOGIA

4.1 Procedimentos metodológicos

Este trabalho é caracterizado como uma pesquisa aplicada ou tecnológica, pois seus objetivos estão atrelados ao desenvolvimento de conhecimentos científicos sobre *machine learning* para aplicação em problemas específicos de gestão de riscos no mercado de crédito. Assim, as abordagens executadas caracterizam-se como quantitativa e descritiva, pois ao conjunto de dados utilizado buscou-se descrever seus atributos, analisar o comportamento do fenômeno de interesse e aplicá-los em técnicas estatísticas. Quanto aos seus objetivos, o estudo em questão pode ser considerado como exploratório, tendo em vista que sua intenção é proporcionar mais familiaridade sobre o tema proposto e seus fenômenos relacionados (FONTELLES, 2009).

Com intuito de desenvolver as análises, coletou-se os dados disponíveis no trabalho de Li, Mickel e Taylor (2018), e por meio de uma amostra estratificada pela variável "***MIS_STATUS***" (que indica se a empresa está ou não inadimplente), extraiu-se 30 mil registros para composição do conjunto de dados. Na Figura 1 pode-se observar as etapas executadas no desenvolvimento dos resultados. A base de dados foi submetida a revisão, tratamento, procedimentos de avaliação de valores ausentes e transformações de variáveis na etapa de pré-processamento. Em seguida, as variáveis foram resumidas e apresentadas no estudo descritivo. Posteriormente, os dados foram particionados em 70% destinados ao treinamento dos modelos e 30% para a fase de teste. Visando contornar o desbalanceamento das classes da variável resposta "***MIS_STATUS***", foram aplicadas as técnicas de *Undersampling* e *OverSampling* na etapa de treinamento dos modelos, resultando assim na possibilidade de treinamento dos algoritmos em três cenários: 1) desbalanceado: cenário real sem aplicação de técnica de balanceamento no treinamento; 2) cenário *Undersampling* e 3) cenário *Oversampling.*

Com base nos cenários indicados, foram elaborados modelos utilizando os algoritmos regressão logística, *Naive bayes, decision tree* e *random forest*. Aspirando a avaliação e melhoria do treinamento dos modelos, foi utilizado o *K-fold cross validation* com *K* igual a 5 partições. Após o treinamento, os modelos foram aplicados na classificação dos registros reservados na base de teste, a fim de avaliar o poder de generalização. As medidas usadas para verificar a performance preditiva foram a acurácia, sensibilidade, especificidade e precisão. Já para a comparação e seleção do melhor modelo, a métrica

utilizada foi a AUC, apresentada em conjunto com as curvas ROC dos modelos. Por fim, todas as análises e gráficos foram feitos utilizando a linguagem de programação R sob o ambiente de desenvolvimento integrado R Studio, versão 4.1.2.

Figura 1 - Fluxograma dos procedimentos metodológicos executados no presente estudo

Fonte: Elaborado pelo autor

Nesse sentido, os tópicos seguintes deste capítulo abordam a teoria e explanam os procedimentos e métodos aplicados desde a coleta dos dados, configurações dos algoritmos, aplicação e avaliação dos modelos.

4.2 Coleta dos dados

O conjunto de dados utilizado trata-se de uma amostra estratificada de 30 mil registros, segmentada pela variável ***"MIS_STATUS"***, extraídos de um *dataset* da *US Small Business Administration* (SBA) disponibilizado no estudo de Li, Mickel e Taylor (2018). Os dados representam as solicitações de empréstimos em bancos realizadas por pequenas empresas dos Estados Unidos, sendo que um percentual do valor concedido é garantido pela entidade SBA.

A SBA foi fundada em 1953 com o objetivo de auxiliar as pequenas empresas no mercado de crédito dos Estados unidos, e um de seus métodos de apoio refere-se ao programa de garantia de crédito. Nesta ação a SBA atua como uma provedora de seguros, com o objetivo de mitigar os riscos para as instituições financeiras, isto significa que, se

uma empresa tomadora de empréstimo não honrar com o compromisso de pagamento, a SBA cobre o montante que garantiu no momento inicial (LI, MICKEL e TAYLOR, 2018).

No Quadro 1 pode-se observar as variáveis utilizadas no desenvolvimento das análises e composição dos modelos. Para este trabalho, a variável resposta corresponde a ***"MIS_STATUS",*** que indica se a empresa pagou integralmente o débito **(PIF)** ou não honrou com o compromisso de pagamento **(GHOFF)**.

Quadro 1 – Descrição das variáveis presentes no estudo Li, Mickel e Taylor (2018) e utilizadas neste trabalho

Variável	**Tipo**	**Descrição da variável**
MIS_Status	Qualitativa	PIF= Pago integralmente e CHGOFF= Status do empréstimo baixado.
NAICS	Qualitativa	Código identificador do setor da indústria norte-americana
Term	Quantitativa	Prazo do empréstimo em meses
NoEmp	Quantitativa	Número de funcionários da empresa
NewExist	Qualitativa	1 = Negócio preexistente e 2 = Novo negócio
CreateJob	Quantitativa	Número de empregos criados
RetainedJob	Quantitativa	Número de empregos retidos
UrbanRural	Qualitativa	0 = Indefinido, 1 = Urbano e 2 = Rural.
LowDoc	Qualitativa	Programa de Empréstimo LowDoc: Y = Sim e N = Não.
GrAppv	Quantitativa	Valor bruto do empréstimo aprovado pelo banco
SBA_Appv	Quantitativa	Valor garantido do empréstimo aprovado pela SBA

Fonte: Elaborado pelo autor

4.3 Preparação dos dados

Na área de *machine learning* existe algumas etapas que são frequentemente executadas, identificadas como *data science process*, e podem ser resumidas em quatro fases: 1) Definição do problema; 2) Tratamento dos dados; 3) Treinamento dos modelos; e 4) Avaliação das performances (CORDEIRO, 2020). O tratamento dos dados é considerado um dos principais passos para o sucesso no desenvolvimento de modelos com bom desempenho preditivo. Santos et al. (2019) apontam que, de modo geral, o pré-processamento está ligado a seleção das variáveis, exclusão de observações faltantes ou utilização de técnicas para imputação de informação e transformação dos dados.

A verificação de ocorrência e tratamento de dados ausentes (*missing values)* é de suma relevância no pré-processamento, pois além da maioria dos algoritmos não trabalharem com registros faltantes, os resultados obtidos podem sofrer distorções se o tratamento escolhido não for cuidadosamente pensado (BATISTA, 2003; CORDEIRO,

2020). Batista (2003) apresenta em seu estudo algumas técnicas que podem ser utilizadas para a correção de *missing values*, sendo elas:

- **Exclusão dos registros com informações faltantes**: Consiste em remover o registro por completo da base de dados.
- **Imputação por algum valor constante**: Refere-se à ação de substituir o valor faltante por um valor fixo. Ex.: 0.
- **Imputação por estatísticas da variável**: Esse método propõe que os valores faltantes sejam substituídos pela média ou mediana quando as variáveis forem quantitativas e pela moda quando qualitativa.
- **Desenvolvimento de modelos preditivos para previsão de informações faltantes:** É possível desenvolver modelos preditivos para preencher as informações faltantes por estimativas, de modo que a variável com dados ausentes é estimada com base nas demais.

Após ter solucionado a presença de *missings values*, é preciso averiguar a necessidade de submeter as variáveis a alguma transformação para melhorar o desempenho dos algoritmos. As transformações mais utilizadas estão relacionadas a normalização de variáveis quantitativas, que tem como objetivo transformar os valores de seus intervalos originais para intervalos específicos, como por exemplo, o intervalo [0,1] (BATISTA, 2003), seja pelas transformações de Box-Cox, ou de casos particulares como a aplicação da escala logarítmica, cúbica, linear e entre outras (PINO, 2014). Para a utilização de alguns algoritmos é preciso que as variáveis qualitativas se tornem quantitativas, seja por auxílio de variáveis *dummies* ou de classes ordinais a depender da tipologia do atributo (BATISTA, 2003).

No presente estudo foi aplicado o log nas variáveis quantitativas, com objetivo de reduzir a assimetria dos dados. A variável dependente "***MIS_STATUS***" foi transformada em um fator que indica se as empresas pagaram ou não o empréstimo recebido, de modo que a classe **(PIF)** tornou-se **(Sim)** e a classe **(GHOFF)** foi transformada em **(Não)**. A variável "***NAICS***" continha 20 classes que segmentavam as empresas pelo ramo de atuação no mercado. Deste modo, com o objetivo de reduzir o número de segmentações, buscou-se reagrupar as empresas em apenas quatro grupos, como mostra a Tabela 1.

Tabela 1 – Novas classes criadas pelo agrupamento dos ramos de atuação da variável NAICS

ID	Ramo	Ramo agrupado	Classe
42	Comércio por atacado	Comércio	C
44–45	Comercio de varejo	Comércio	C
48–49	Transporte e armazenamento	Comércio	C
55	Gestão de empresas e empreendimentos	Comércio	C
11	Agricultura, silvicultura, pesca e caça	Indústria e Agropecuária	I
21	Mineração, pedreiras e extração de petróleo e gás	Indústria e Agropecuária	I
23	Construção	Indústria e Agropecuária	I
31–33	Fabricação	Indústria e Agropecuária	I
51	Em formação	Outros	O
81	Outros serviços (exceto administração pública)	Outros	O
22	Serviços de utilidade pública	Serviços	S
52	Finanças e seguros	Serviços	S
53	Imóveis e aluguel e leasing	Serviços	S
54	Serviços profissionais, científicos e técnicos	Serviços	S
56	Serviços administrativos e remediação de resíduos	Serviços	S
61	Serviços educacionais	Serviços	S
62	Cuidados de saúde e assistência social	Serviços	S
71	Artes, entretenimento e recreação	Serviços	S
72	Serviços de hospedagem e alimentação	Serviços	S
92	Administração pública	Serviços	S

Fonte: Elaborado pelo autor

4.4 Particionamento dos dados

Um dos principais objetivos em *Machine Learning* é alcançar uma boa capacidade de generalização, ou seja, construir um modelo a partir de dados conhecidos (etapa de treinamento) que seja capaz de descrever com precisão o comportamento de dados desconhecidos (etapa de teste). Diante desse objetivo, a base de dados é particionada em dois grupos, nomeados de treino e teste. Na fase de treinamento, os registros são utilizados para escolha dos parâmetros, comparação e seleção do modelo final. Após ter escolhido o modelo final, os dados reservados no conjunto de teste são utilizados para verificar seu poder de generalização, ou seja, o modelo produz estimativas com base em dados que não fizeram parte do ajuste dos parâmetros (BOEHMKE e GREENWELL; 2019).

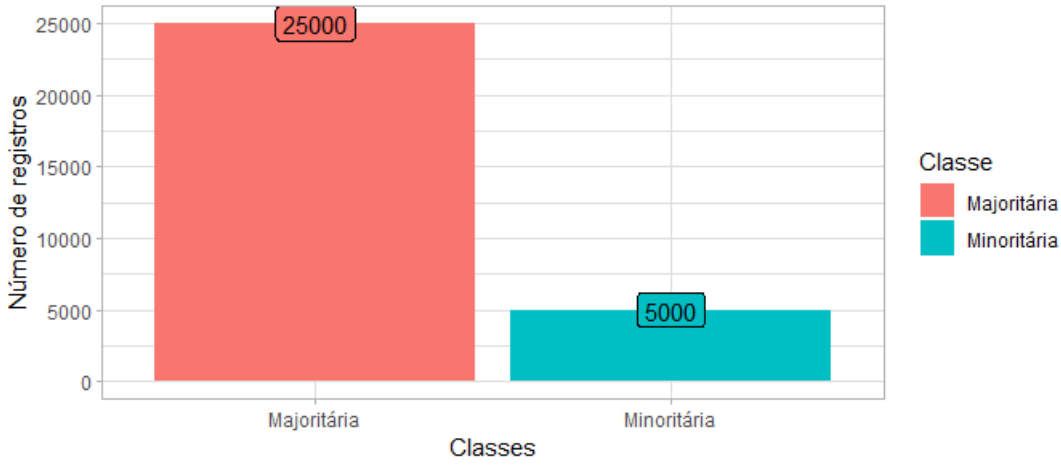
Gallicchio et al. (2017) indicam que a etapa de design do particionamento dos dados pode ser um tema ainda mais explorado pelos pesquisadores. Boehmke e Greenwell (2019) apontam que as recomendações típicas para dividir os dados em conjunto de treino e teste estão entre (60%-40%), (70%- 30%) ou (80%-20%). Neste estudo, 70% dos dados foram destinados à fase de treinamento e os 30% restantes foram reservados para a fase de teste,

estas proporções na divisão estão em concordância com a indicação de Gallicchio et al. (2017), e foi também utilizada pelos estudiosos Aniceto (2016) e Cordeiro (2020).

4.5 Dados desbalanceados

Um conjunto de dados é considerado desbalanceado quando a variável resposta do problema a ser modelado tem muito mais registros em uma determinada classe do que em outras. A classe com maior número de registros é denominada como majoritária, enquanto as demais são chamadas de classes minoritárias (CORDEIRO, 2020). Como exemplo, supondo uma base de dados que contém 30 mil registros e que a variável resposta tem a distribuição apresentada na Figura 2, pode-se observar um nível de desbalanceamento, pois uma classe está concentrando 83% dos registros.

Figura 2 – Representação do desbalanceamento dos dados de acordo com as classes da variável resposta (Cenário hipotético)



Fonte: Elaborado pelo autor

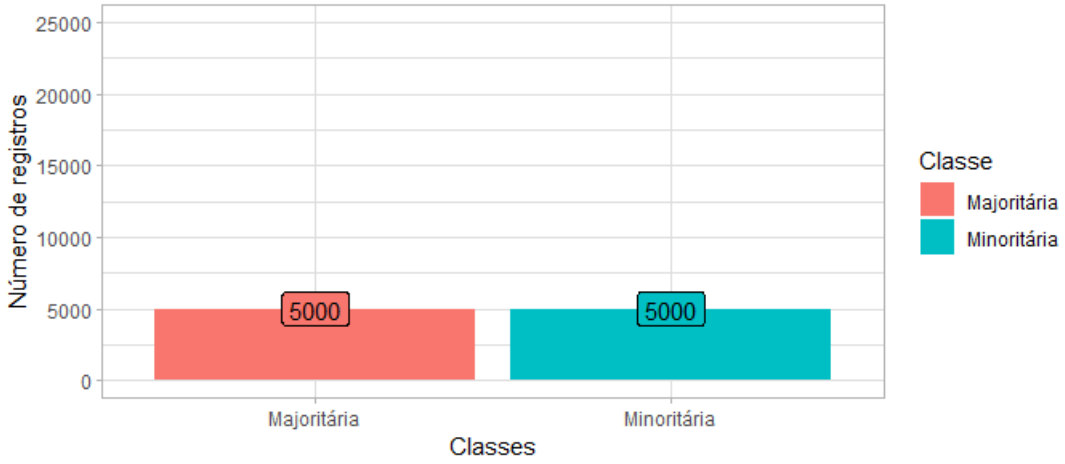
Grande parte dos algoritmos de *machine learning* produzem a ilusão de um bom desempenho em problemas de classificação com dados desbalanceados, pois muitos foram projetados para lidar com número de registros iguais em cada classe. Quando há o desbalanceamento, os algoritmos tendem a aprender que os poucos exemplos da classe minoritária não são importantes e podem ser ignorados, a fim de se obter um bom desempenho, portanto acabam acertando a classe majoritária, não pela capacidade do método identificá-la, mas sim por ser muito mais frequente (BRANCO et al., 2016).

Stelzer (2019) aborda em seu estudo diversas técnicas que buscam contornar o problema do desbalanceamento dos dados. Essas técnicas podem ser divididas em duas

principais abordagens: pré-processamento e algorítmica (BARELLA, 2015). Entre as abordagens de pré-processamento tem-se os métodos de reamostragem, que visam o balanceamento dos dados no treinamento através da redução da influência da classe majoritária, podendo ser classificadas em: (1) mecanismos que reduzem a classe majoritária (*Undersampling*); (2) técnicas que maximizam a participação da classe minoritária (*Oversampling*) e (3) técnicas que combinem as estratégias anteriores (CORDEIRO, 2020; PIEDADE, 2020).

O método de *Undersampling* é baseado em subamostragem para a remoção de elementos da classe majoritária, essa remoção pode ser feita de maneira aleatória (subamostragem aleatória) ou por algum critério de seleção (subamostragem informativa). O número de elementos a ser retirados pode variar, porém conforme apresentado na Figura 3, espera-se que ao final do processo a proporção da classe majoritária seja igual ao da classe minoritária (BARELLA, 2015). Essa técnica mitiga o efeito de *overfitting*, contudo sua utilização excessiva pode causar a perda de informações úteis, afetando assim o desempenho do modelo (LIU et al., 2019).

Figura 3 - Representação do balanceamento dos dados usando a técnica de *Undersampling* (Cenário hipotético)



Fonte: Elaborado pelo autor

Já o método de *Oversampling* tem como objetivo maximizar a classe minoritária. Seja pela replicação de registros (sobreamostragem com repetição), isto é, um subconjunto da classe inferior é selecionado e replicado, ou através da geração de dados artificiais. Assim, semelhante ao método apresentado anteriormente, percebe-se na Figura 4 que o processo é repetido até que se tenha o número de registros iguais em ambas as classes

(BARELLA, 2015). Vale salientar que sua aplicação pode gerar *overfitting* se as amostras forem muito semelhantes (LIU et al., 2019).

Figura 4 - Representação do balanceamento dos dados usando a técnica de *Oversampling* (Cenário hipotético)

Fonte: Elaborado pelo autor

Neste estudo, a etapa de amostragem dos dados para o desenvolvimento dos modelos seguiu três situações. No primeiro cenário, considerado desbalanceado, as proporções dos dados no treinamento e teste seguiram o exposto na seção 4.4. Ademais, tanto para a base de treinamento quanto para a base de teste, foi utilizado a amostragem estratificada pela proporção das classes da variável resposta "***MIS_STATUS***", de modo que o percentual de clientes adimplentes e inadimplentes sejam semelhantes em ambas as bases. Vale salientar que os demais cenários advindos das aplicações das técnicas *Undersampling e Oversampling***,** foram realizados apenas no conjunto de treinamento, evitando assim modificações da proporção real da variável resposta na base de teste.

4.6 Regressão logística múltipla

A regressão logística é um modelo linear generalizado que visa explicar a relação de uma variável dependente binária com uma ou mais variáveis preditoras. Supondo que n seja o número de variáveis independentes representadas pelo vetor $X(x_{1,}\, x_{2,}\, \dots, x_{n,})$ e Y seja uma variável aleatória binária (FIGUEIRA, 2006; PAIVA, 2015), temos que o modelo de regressão logística com n variáveis independentes será

$$E(Y_i \mid x_i) = = \frac{\exp\,(\,\beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \cdots + \beta_n\, x_n)}{1 + \exp\,(\,\beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \cdots + \beta_n\, x_n)} \qquad (1)$$

onde,

$E(Y_i \mid x_i)$ é o valor esperado de Y dado o valor de x;

Y_i é uma variável binária, isto é, assume apenas dois valores (0 ou 1);

β_0, β_1, … β_n são os parâmetros do modelo estimados pelo método da máxima verossimilhança.

Simplificando a notação, temos que $p(x_i) = E(Y_i \mid x_i)$. Assim, $p(x_i)$ pode ser transformada na forma linear pela função logit(). Resultando a função de ligação $g(x_i)$, tal que

$$g(x_i) = ln\left[\frac{p(x_i)}{1- p(x_i)}\right] \quad (2)$$

$$g(x_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (3)$$

Logo, $p(x_i)$ é determinado por

$$p(x_i) = \left[\frac{exp(g(x_i))}{1+ \exp(g(x_i))}\right] + \varepsilon_i \quad (4)$$

Deste modo, a variável dependente Y_i pode ser representada por

$$Y_i = p(x_i) + \varepsilon_i \quad (5)$$

onde o termo ε_i é um erro aleatório e representa a diferença entre o valor observado de Y_i e o valor estimado para $\hat{Y}_i$ dado x_i. Vale salientar que enquanto no modelo de regressão linear um dos pressupostos é que o erro tenha distribuição normal, na regressão logística os erros seguem distribuição de Bernoulli com média 0 e variância $p(x_i) - [1 - p(x_i)]$ (PAIVA, 2015).

As estimativas dos parâmetros (β_0, β_1, … β_n) são calculadas pelo método da máxima verossimilhança. Sabendo que a probabilidade de $Y_i = 1$ é igual a $P(Y_i = 1 \mid x_i) = p(x_i)$ e que para de $Y_i = 0$ é igual a $P(Y_i = 1 \mid x_i) = 1 - p(x_i)$. Forti (2018) explica que uma maneira de expressar matematicamente a contribuição de cada indivíduo na função de verossimilhança é

$$L_i = p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i} \quad (6)$$

donde y_i assume valores 0 ou 1 e o índice i varia entre os números positivos. Ademais, como as observações são independentes, a função de verossimilhança pode ser escrita como

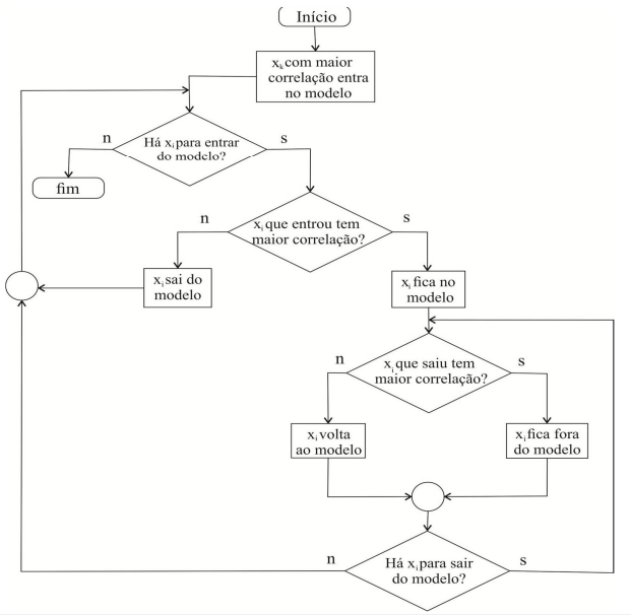
$$L_i = \prod_{b=1}^{B} p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i} \quad (7)$$

Baseado no princípio da parcimônia, um bom modelo é definido pela sua capacidade de explicação da variável resposta com o mínimo possível de variáveis explicativas (OLIVEIRA, 2020). Diante desse contexto, a utilização de métodos para seleção das variáveis que devem constituir o modelo torna-se importante, pois muitas variáveis podem não contribuir na discriminação (FORTI, 2018). Segundo Paiva (2015), as técnicas mais utilizadas para seleção de variáveis na literatura são:

a) ***Backward-*** Parte de um modelo inicial composto por todas as variáveis e, a cada passo dado, as variáveis que menos contribuem para o modelo vão sendo retiradas até que se consiga o melhor modelo.
b) ***Forward -*** É iniciado com um modelo sem nenhuma variável preditora, e a cada passo, são incluídas aquelas que mais contribuem para o modelo, até que seja identificado o melhor modelo.
c) ***Stepwise-*** É uma combinação dos dois métodos anteriores, pois inicia-se sem nenhuma variável, e a cada etapa de inclusão também é feita a exclusão de variáveis.

No presente estudo foi utilizado o método de *Stepwise*, apresentado na Figura 5. Sua utilização garante uma forma ágil e eficiente de avaliar muitas variáveis, e concomitantemente, analisar diversos modelos gerados por suas combinações (PAIVA, 2015).

Figura 5 – Fluxograma para seleção de variáveis utilizando o método de *Stepwise*



Fonte: Alves, Lotufo e Lopes (2013)

4.7 Naive Bayes

O advento do raciocínio bayesiano é dado pela inferência probabilística. Sua fundamentação provém de que as quantidades de interesse estão atreladas a uma probabilidade de distribuição e que as decisões podem ser tomadas pela análise dessas probabilidades em um conjunto de dados (ANDRADE, 2008). Uma rede bayesiana define-se como uma representação compacta de uma tabela de probabilidades do universo, apresentando de maneira simples as relações de casualidade das variáveis envolvidas (MARQUES, 2002). Em uma rede Bayesiana a distribuição conjunta de probabilidades de um número de variáveis discretas $\{x_1, x_2, x_3, \dots, x_n\}$ é dada pela regra da cadeia

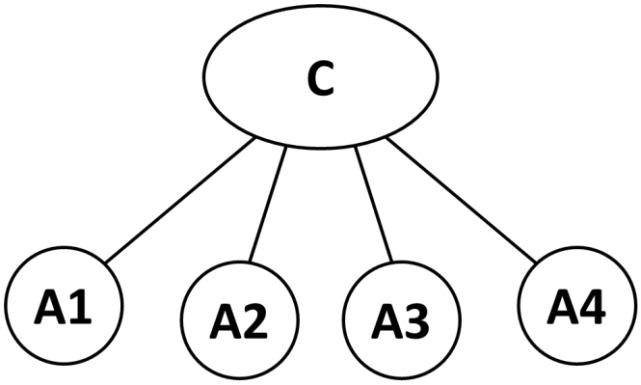
$$P(x_1, x_2, x_3, \dots, x_n) = \prod_{i=1}^{n} P(X_i \mid Pa_i) \qquad (8)$$

Definindo $\{\theta_i = P(X_i \mid Pa_i), i = 1, \dots, n\}$ como o parâmetro de uma rede bayesiana, em que θ_i é uma tabela condicional da probabilidade de X_i dado Pa_i. Assim, uma rede Bayesiana é dada por um vetor $\theta_t = \{\theta_1, \theta_2, \theta_3, \dots, \theta_n\}$ que representa um conjunto de tabelas condicionais de $\{x_1, x_2, x_3, \dots, x_n\}$ dado Pa_i (NEAPOLITAN, 2007).

As redes bayesianas podem ser aplicadas em problemas que envolvam classificação, resultando nos classificadores bayesianos, sendo os mais simples chamados de *Naive Bayes*. Estes classificadores assumem que a presença ou ausência de uma

determinada característica está relacionada a outros elementos (SANTOS, 2013), conforme apresentado na Figura 6.

Figura 6 - Estrutura de um classificador de *Naive Bayes*



Fonte: Elaborado pelo autor

Dado um conjunto das variáveis $A_i = \{A_1, A_{2\,,}\, A_3, \dots, A_n, C\}$, em que C é a variável dependente e as demais são atributos preditores independentes, Kracher (2009) apresenta que a distribuição conjunta de probabilidades advinda do classificador *Naive Bayes* é dada por

$$P(A_1, A_{2\,,}\, A_3, \dots, A_n, C) = P(C) \prod_{i=1}^{n} P(A_i \mid C) \qquad (9)$$

Considerando um classificador bayesiano com atributos discretos e que a variável dependente é uma classe C com valores {0,1}, a probabilidade de classificação de um novo caso A_i, em C =1, é dada por:

$$P(C = 1 \mid\ A_i = a_i\ ) = \frac{P(C = 1)\,.\,P(A_1 = a_{1\,,} \dots, A_n = a_n\ \mid C = 1)}{P(A_1 = a_{1\,,} \dots, A_n = a_n)} \qquad (10)$$

De maneira similar apresentada na equação (10) pode-se calcular a probabilidade de classificação de um novo caso dado que a classe seja C=0. Assim, tem-se que uma nova observação é classificada como C=1, se e somente se

$$\frac{P\big(C = 1 \mid\ A_1 = a_{1\,,} \dots, A_n = a_n\ \big)}{P\big(C = 0 \mid\ A_1 = a_{1\,,} \dots, A_n = a_n\ \big)} \geq 1 \qquad (11)$$

A equação (11) pode ser reescrita como

$$\frac{P(C=1)\,.P(A_1=a_1,\ldots,A_n=a_n\mid C=1)}{P(C=0)\,.\;A_1=a_1,\ldots,A_n=a_n\mid C=0)}\geq 1 \qquad (12)$$

Diante desse contexto, Kracher (2009) e Santos (2013) apresentam em seus estudos que no caso do classificador *Naive Bayes*, o critério para um novo caso A_i ser classificado como C =1 é dado por

$$\frac{P(C=1)}{P(C=0)}\prod_{i=1}^{n}\frac{P(A_i=a_i\mid C=1)}{P(A_i=a_i\mid C=0)}\geq 1 \qquad (13)$$

Lobato e Carvalho (2021) indicam que se alguma variável do vetor A_i for quantitativa, então é comum supor a associação das classes com uma distribuição Gaussiana, segmentando as classes e calculando a média e variância da variável x. Assim, a probabilidade é dada por

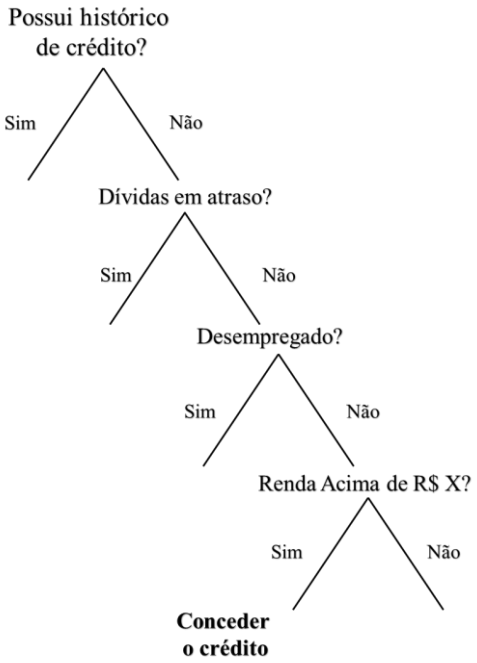
$$P(X=x\mid\ C=c)=\frac{1}{\sqrt{2\pi\sigma_c^2}}\,e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}} \qquad (14)$$

Webb e Keogh (2010) resumem o *Naive Bayes* como o algoritmo de aprendizado que se baseia na regra de *bayes* e na forte suposição de que as variáveis são independentes dada a classe. Segundo os autores, a sua eficiência computacional, facilidade de compreensão e muitas outras características fazem com que este algoritmo seja amplamente utilizado.

4.8 Decision tree

A técnica de *Decision Tree* possui uma estrutura de árvore e tem como objetivo segmentar os resultados hierarquicamente, isto é, há uma priorização. O modelo é desenvolvido por uma série lógica, os atributos mais importantes são apresentados na árvore como os primeiros nós e as variáveis menos relevantes são apresentadas nos nós subsequentes. Deste modo, ao analisar a árvore gerada, é possível identificar os fatores mais importantes para explicação da variável resposta (LEMOS, STEINER e NIEVOLA, 2005). Na Figura 7 é possível observar um exemplo de estrutura de um modelo baseado no algoritmo de *Decision tree*.

Figura 7- Exemplo de estrutura de uma árvore de decisão incompleta para concessão de crédito



Fonte: Elaborado pelo autor

Existem diversos algoritmos para o desenvolvimento de árvores de decisão, e a escolha do qual será utilizado dependerá das variáveis envolvidas na modelagem (LEMOS, STEINER e NIEVOLA, 2005). Segundo Borba (2021), entre os mais utilizados pode-se citar:

- *Iterative Dichotomiser* 3 (ID3): Sua aplicação é indicada quando os atributos são categóricos e a variável dependente é binária.
- *Chi-Squared Automatic Interaction Detector* (CHAID): Indicado quando todas as variáveis são categóricas.
- *Classification and regression Tree* (CART ou C&RT): Não possui restrições quanto ao tipo de variável.

Neste estudo, o algoritmo utilizado para o desenvolvimento de árvores é baseado no CART. Em uma árvore de decisão busca-se prever qual partição cada registro pertence com base nas observações de treinamento mais comuns na região a que está inserido (LOBATO e CARVALHO, 2021). O conjunto de possíveis valores para a partição é calculado pelo ponto médio de cada unidade de respostas distintas ao longo de cada atributo. Isto é, para cada p respostas distintas tem-se p_i possíveis valores para partição observada (MYLES et al., 2004; ANICETO, 2016; LOBATO e CARVALHO, 2021).

Assim, MYLES et al. (2004) apresentaram que a probabilidade de um caso aleatório pertencer a classe (j) é dado por

$$P_i = \frac{N_j(t)}{N(t)} \tag{15}$$

onde,

$N_j(t)$ é o número de registros pertencentes à classe j;

$N(t)$ é o número de amostras em t nós.

Entretanto, o P_i não é uma medida tão sensível para o cultivo da árvore (LOBATO e CARVALHO, 2021) e a literatura sugere a análise de mais duas medidas, sendo uma delas chamada de entropia, calculada por

$$Ent(y) = -\sum \left(\frac{N_j(t)}{N(t)}\right) log_2 \left(\frac{N_j(t)}{N(t)}\right) \tag{16}$$

A outra medida avaliada é o índice de Gini (MYLES et al., 2004) e seu cálculo é dado por

$$\mathcal{G}(y) = \mathcal{G}(p_1, p_2, \dots, p_j) = \sum_j p_j \sum_{i \neq j} p_i$$

$$= \sum_j p_j \ (1 - p_j)$$

$$\mathcal{G}(y) = 1 - \sum_j p_j^2 \tag{17}$$

Andrade (2008) salienta que os critérios avaliados acima apenas medem o impacto do particionamento, e que não indicam quais e nem onde as variáveis devem ser particionadas. Deste modo, o ganho de informação com a partição pode ser calculado da seguinte maneira

$$\Delta_i = Ent(y) - p_d Ent(y_d) - p_e Ent(y_e) \tag{18}$$

ou

$$\Delta_i = \mathcal{G}(y) - p_d \mathcal{G}(y_d) - p_e \mathcal{G}(y_e) \tag{19}$$

em que (d) e (e) representam respectivamente direita e esquerda do novo nodo. Destarte, o ganho de informação representa a diferença entre a impureza inicial e a média da impureza com a criação dos novos nodos (ANDRADE, 2008).

Na visão de Lemos, Steiner e Nievola (2005) as vantagens de utilizar o algoritmo *decision tree* estão na facilidade de interpretação, ausência de pressupostos sobre a distribuição particular para os dados, a aderência para qualquer tipo de atributo (quantitativo ou qualitativo) e sua aplicabilidade em qualquer tipo de problema, dado que tenha o número suficiente de registros para treinamento.

4.9 Random Forest

O algoritmo *Random forest* foi proposto por Breiman, seu método constitui um conjunto de árvores de decisão $\{h(x, \theta_k), k = 1,2, \dots, n\}$, onde θ_k são vetores randomizados distribuídos identicamente e x é o valor de entrada. Sua ideia principal é a redução da correlação de um conjunto de árvores de decisão com baixo poder preditivo, até que se tenha um modelo com maior acurácia, sem aumentar muito a variância (ANICETO, 2016; FORTI, 2018; LAKSHMI e KAVILLA, 2018). Dado que as árvores ajustadas têm baixa correlação umas com as outras, a previsão é dada com base na união dos ajustes. Mesmo havendo árvores com baixo poder preditivo, a precisão das outras pode produzir uma previsão razoável (LI et al., 2020).

Forti (2018) explica que a redução da correlação das árvores é ocasionada tanto pela reamostragem ocorrida nos registros, quanto pela reamostragem das variáveis, essa redução pode ser representada matematicamente, como

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \qquad (20)$$

em que

ρ é número de parâmetros;

σ^2 é a variância;

B refere-se ao número de árvores.

Como as árvores são identicamente distribuídas, espera-se que a média de B seja a mesma para qualquer uma delas. Isto é, o viés do conjunto de árvores não correlacionadas é igual ao das árvores individuais, implicando que a melhoria esperada seja ocasionada

pela redução da variância. Assim, conforme apresentado na equação (20), à medida que B aumenta o segundo termo da soma é diminuído, resultando na redução da correlação entre as árvores sem aumentar muito a variância (HASTIE, TIBSHIRANI e FRIEDMAN, 2009).

A randomização ocorrida é limitada a um subconjunto aleatório $mtry$ (Parâmetro de aleatoriedade) dos p recursos originais (GISLASON, BENEDIKTSSON e SVEINSSON, 2006; CORDEIRO, 2020). Tendo em vista que o problema proposto se refere a uma classificação, então o $mtry$ será dado por $\sqrt{p}$ (HASTIE, TIBSHIRANI e FRIEDMAN, 2009). Deste modo, no Quadro 2 é formalmente apresentado a estrutura deste algoritmo RF.

Quadro 2- Estrutura do Algoritmo *Random forest* para classificação

Algoritmo do Random Forest para classificação

- Para i = 1 até B, faça:
  - Gere uma amostra de bootstrap dos dados originais;
  - Construa uma árvore de regressão/classificação para os dados iniciais;
  - Para cada divisão, faça:
    - Selecione $mtry$ variáveis aleatoriamente de todas as variáveis p;
    - Escolha a melhor variável/ponto de divisão entre os $mtry$;
    - Divida o nó em dois nós;
  - Use critérios de parada de um modelo de árvore típico para determinar quando uma árvore estará completa (sem poda);
  - Conjunto de árvores de saída $\{T_b\}_1^B$.

A previsão para um novo ponto x é dado por:

Seja $\hat{C}_b(x)$ a previsão da b-ésima floresta aleatória da árvore, então a previsão $\hat{C}_{rf}^B(x) =$ maior número de classificações $\left\{\hat{C}_b(x)\right\}_1^B$.

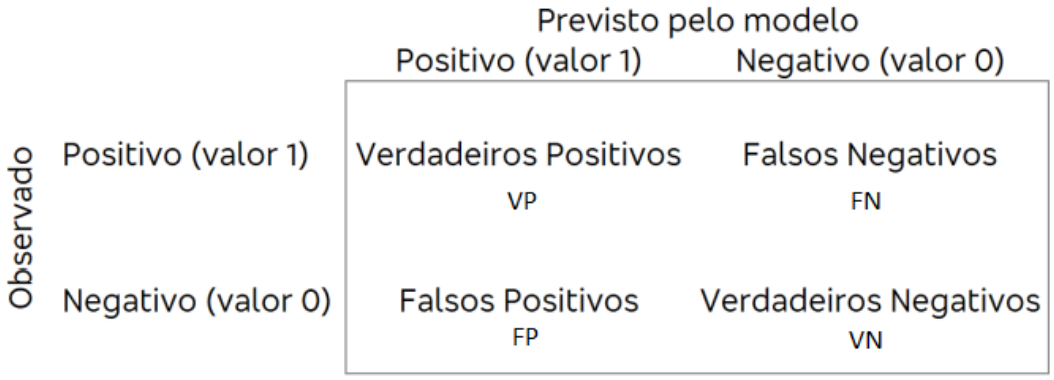
Fonte: Adaptado de Hastie, Tibshirani e Friedman (2009) e Forti (2018)

4.10 Métricas de Avaliação

Uma vez que os modelos são aplicados aos dados, seu desempenho preditivo deve ser avaliado empregando algumas métricas (STELZER, 2019). A matriz de confusão é uma maneira fácil de observar se o modelo está performando bem, isto é, se está conseguindo prever adequadamente as classes positivas e negativas (KARCHER, 2009). Na matriz de confusão existem duas dimensões: uma destinada às classes observadas do conjunto de dados e outra destinada as classes preditas pelo modelo. As estimativas de classificação corretas, em comparação com os valores reais, encontram-se na diagonal principal da matriz e são apresentadas na Figura 8 pelas classes (VP + VN), enquanto as predições erradas são apresentadas na diagonal secundária (FP + FN) (SANTOS, 2013).

As estimativas dos modelos são apresentadas de forma binária, e para auxiliar na dicotomização das previsões faz-se necessário a definição de um ponto de corte p, e comparar cada probabilidade estimada $P(\hat{y})$ com o ponto definido. Se a $P(\hat{y}) \geq p$, então assume-se que o resultado predito para a variável resposta deverá ser igual a 1, caso contrário, deverá ser igual a 0. Geralmente, o valor do ponto de corte utilizado é 0,5 (PAIVA, 2015). Como exemplo, supondo que a estimativa para variável resposta $\hat{y}$ seja positiva quando recebe 1 e negativa quando recebe zero. Se a $P(\hat{y}\,)$ = 0,67, então considerando o ponto de corte $p = 0{,}5$, tem-se que $\hat{y}$ receberá a classe positiva, pois $P(\hat{y}\,) \geq p$.

Figura 8 - Modelo de uma matriz de confusão binária



Fonte: Adaptado de Cordeiro (2020)

A medida de acurácia representa o quanto o modelo foi capaz de capturar os verdadeiros positivos e verdadeiros negativos, e representá-los como uma proporção do

total de previsões (LAKSHMI e KAVILLA, 2018; LI et al., 2020), essa medida é calculada por

$$Acurácia = \frac{(VP+VN)}{VP+FN+VN+FP} \quad (21)$$

O *recall* ou sensibilidade trata-se da capacidade de identificação correta dos classificados como positivo, quanto mais próximo de 1 mais sensível será o modelo, implicando numa maior exatidão na identificação de casos positivos (ANDRADE, 2008; LAKSHMI e KAVILLA, 2018; LI et al., 2020), sua representação matemática é dada por

$$Sensibilidade = \frac{VP}{(VP+FN)} \quad (22)$$

Já a especificidade tem como objetivo avaliar qual foi a performance do modelo em detectar os casos negativos, um modelo muito específico aponta os casos negativos com alto grau de acerto (ANDRADE, 2008; LAKSHMI e KAVILLA, 2018; LI et al., 2020). Sendo calculada pela equação

$$Especificidade = \frac{VN}{(VN+FP)} \quad (23)$$

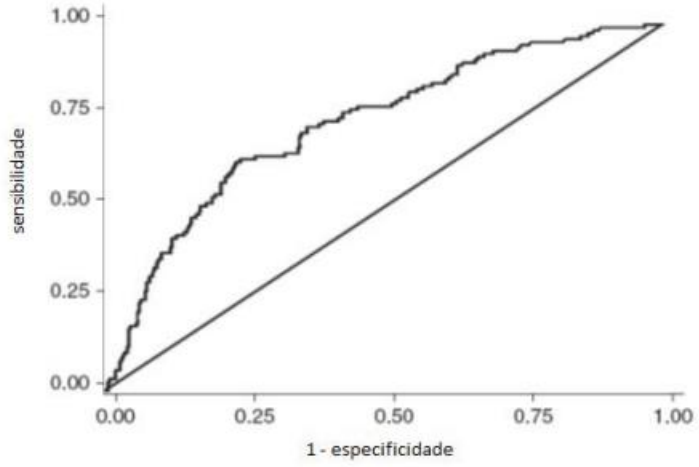
A medida de precisão analisa a quantidade de verdadeiros positivos sobre o número total de casos positivos, conforme apresentado na equação

$$Precisão = \frac{VP}{(VP+FP)} \quad (24)$$

O gráfico da *Receiver Operating Characteristic* (Curva ROC) apresenta o desempenho do modelo em relação sensibilidade e especificidade, conforme apresentado na Figura 9. No eixo x tem-se a taxa de falso positivo (1- especificidade) e no eixo y a taxa de verdadeiro positivo (sensibilidade) para todo ponto de corte (ANDRADE, 2008; CORDEIRO, 2020). A área sob a curva Roc (*Area Under the Curve* – AUC) é calculada para resumir o comportamento apresentado na Figura 9, sendo utilizada na comparação de diferentes modelos, em que ajustes com baixo poder de predição terão um $AUC \leq 0.5$ e bons modelos terão $0.5 < AUC \leq 1$ (SANTOS et al., 2019; PIEDADE, 2020).

À medida que as distribuições de probabilidades estimadas pelo modelo discriminam a variável resposta, a curva ROC aumenta rapidamente até que área sob ela alcance o valor máximo do intervalo (AUC = 1) (PAIVA, 2015). Na prática o modelo que estiver acima é considerado o que obteve a melhor performance, essas métricas variam entre 0,5 e 1, e quanto mais próximo de 1 melhor o desempenho do modelo (ANDRADE, 2008).

Figura 9 - Exemplo do gráfico da curva ROC em vários pontos de corte



Fonte: Paiva (2015)

4.11 Cross-Validation

A *cross-validation* é um eficiente processo para identificação de *overfitting* e para avaliação e treinamento dos modelos. Neste método os dados de treino são repartidos de maneira aleatória em K partes iguais, e o treinamento é feito em $K - 1$ partes, pois uma parte é reservada para o teste (estimação do erro) (YADAV e SHUKA, 2016; CORDEIRO, 2020). O modelo é, então, executado K vezes, sendo que em cada iteração um subconjunto do K é tomando como teste, e os demais destinados ao treinamento, a fim de que seja mensurada a capacidade preditiva do modelo. A iteração só termina na k-ésima rodada, em que todos os subconjuntos de K já foram destinados uma vez a classe de teste (SANTOS, 2013).

Habitualmente o valor de K varia entre 5 e 10 partições, também chamadas de *folds* (ANDRADE, 2008; SANTOS, 2013). Conforme apresentado na Figura 10, no presente

estudo a iteração foi realizada 5 vezes, de modo que em cada iteração uma partição é reservada para a validação, resultando assim um treinamento mais refinado do modelo.

Figura 10- Processo da metodologia *k-fold* do *Cross-validation*

Fonte: Elaborado pelo autor

O valor estimado para taxa de assertividade da classificação será a média da taxa de acertos em cada iteração, conforme apresentado na equação (25). Se dois modelos tiverem acurácias iguais, então deve-se considerar o de menor variabilidade nas classificações (ANDRADE, 2008; SANTOS, 2013).

$$Acurácia\ Final = \frac{\sum_{i=1}^{k} Acurácia_i}{k} \quad (25)$$

Existem mais dois tipos de metodologia do *cross-validation* além da apresentada, a estratificada e a *leave-one-out*. Na estratificada, a proporção de classes do conjunto de dados é levada em consideração na geração dos subconjuntos, ou seja, se nos dados houver duas classes que representam 30% e 60% do total, essas proporções serão seguidas no processo de subamostragem. Já no *leave-one-out*, o valor de k é igual o tamanho da amostra, sendo mais preciso, por aproveitar os dados e não utilizar subamostragem aleatória, contudo, o custo computacional pode afetar a sua utilização (ANDRADE, 2008; SANTOS, 2013; YADAV e SHUKA, 2016).

5 RESULTADOS E DISCUSSÕES

5.1 Pré-processamento dos dados

5.1.1 Dados Ausentes e inconsistentes

Na etapa de pré-processamento dos dados identificou-se que as variáveis "***NewExist***" e "***LowDoc***" continham valores ausentes. Conforme apresentado na Tabela 2, houve 7 registros sem a informação de classificação da empresa "***NewExist***", e no caso da variável que informa a participação no programa de empréstimo "***Lowdoc***", além de 97 registros com *missings*, tem-se que 74 registros estavam com o preenchimento inconsistente, isto é, apresentam valores diferentes do que seria aceito, de acordo com o Quadro 1.

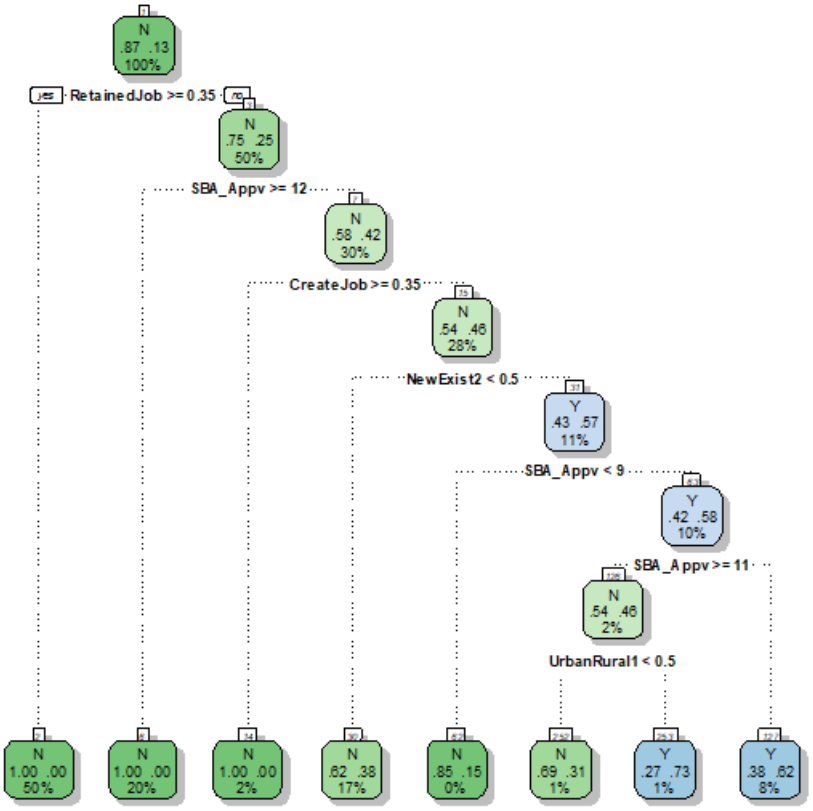
Tabela 2 - Distribuição de valores ausentes ou inconsistentes por variável do conjunto de dados

Variável	***Missing***	**Inconsistência**
NAICS	0	0
Term	0	0
NoEmp	0	0
NewExist	7	0
CreateJob	0	0
RetainedJob	0	0
UrbanRural	0	0
LowDoc	97	74
MIS_Status	0	0
GrAppv	0	0
SBA_Appv	0	0

Fonte: Elaborado pelo autor

Para contornar eventuais impactos negativos nos algoritmos em decorrência dos valores ausentes, os registros com informações faltantes da variável "***NewExist***" foram removidos devido a sua baixíssima representação no conjunto de dados. Já nos os registros com *missing* e inconsistentes da variável "***LowDoc***" buscou-se imputar informações com auxílio das estimativas de um modelo baseado no algoritmo *decision tree*. Neste caso, na Figura 11 tem-se que a variável "***LowDoc***" tornou-se o evento de classificação de interesse (variável dependente) e com base nas demais variáveis exógenas, o modelo determinou se a informação faltante deveria receber a classe de participação (Y) ou não (N) do programa de empréstimo.

Figura 11 - Árvore de decisão do modelo desenvolvido para imputação de dados na variável *Lowdoc*

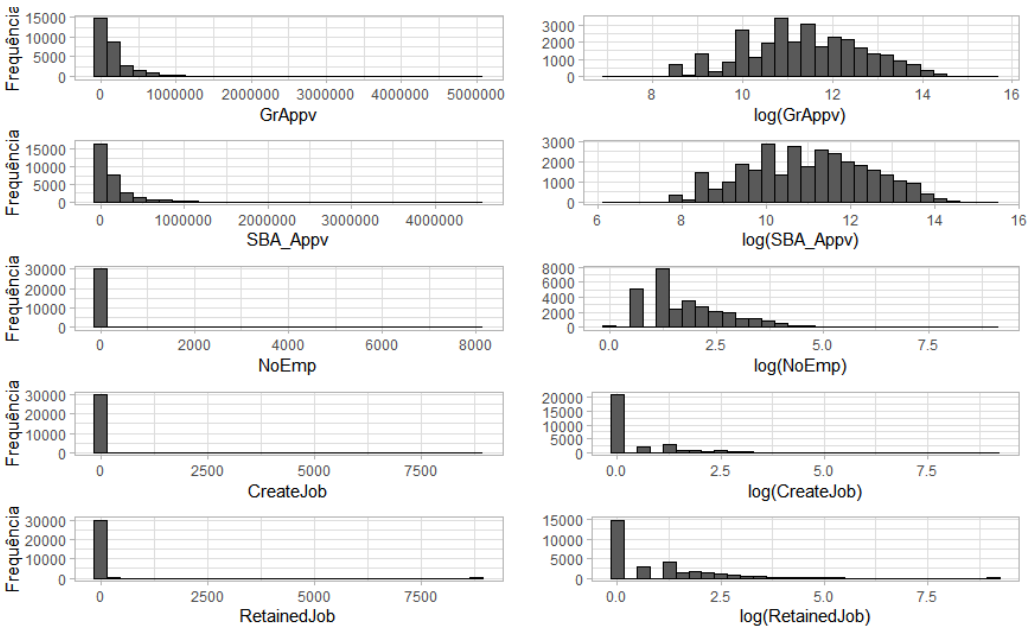


Fonte: Elaborado pelo autor

5.1.2 Transformações dos dados

Nas variáveis quantitativas foi aplicado a função logarítmica, com objetivo de reduzir a assimetria dos dados. Na Figura 12 é possível identificar que a aplicação do log nas variáveis que representam o valor do empréstimo "***GrAppv***" e "***SBA_Appv***" tornou-as aproximadamente simétricas. Por outro lado, não foi suficiente para eliminar a assimetria das variáveis que indicam o número de empregos ("***Noemp***", "***CreatJob***", "***RetainedJob***"). Vale salientar que para a análise descritiva, essas variáveis serão apresentadas na escala original, mas para o ajuste dos modelos, será utilizada a escala logarítmica.

Figura 12 - Distribuição das variáveis (“***GrAppv***”, “***SBA_Appv***”, “***Noemp***”, “***CreatJob***” e “***RetainedJob***”) em escala real e logarítmica

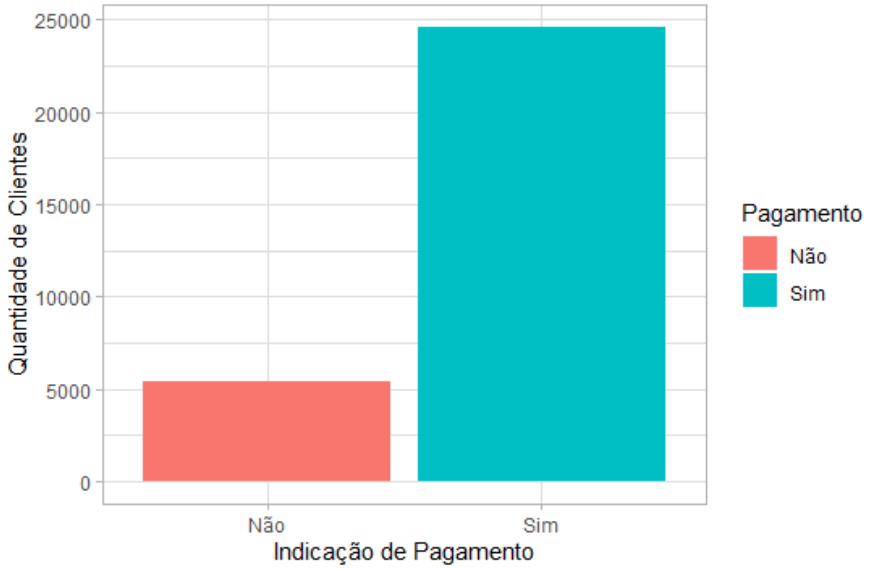


Fonte: Elaborado pelo autor

5.2 Estatística descritiva dos dados

No que se refere à distribuição da variável resposta “***MIS_STATUS***”, percebe-se na Figura 13 que a maioria dos indivíduos honraram com seus compromissos perante as entidades interessadas, constituindo um percentual de 82% de adimplência, enquanto 18% dos clientes foram inadimplentes.

Figura 13 - Distribuição da quantidade de clientes em relação à situação do pagamento do débito perante as instituições financeiras vinculadas a SBA



Fonte: Elaborado pelo autor

Analisando as variáveis explicativas qualitativas em relação à classificação do pagamento, observa-se na Tabela 3 que os empréstimos concedidos para quitação com menos parcelas resultaram um maior percentual de clientes inadimplentes, sendo que 56% dos indivíduos que dividiram o débito em até 30 parcelas não cumpriram com o combinado, já os créditos que tiveram o pagamento dividido em um número superior a 30 parcelas, apresentaram mais de 60% de adimplência, chegando a ultrapassar 90% para os empréstimos divididos em mais de 90 parcelas.

Tabela 3 - Distribuição das variáveis preditoras qualitativas em relação a indicação de pagamento

Definições (Variáveis)	**Indicação de pagamento**	
Número de parcelas (*Term*)	**Não**	**Sim**
(0,30]	1.417 (56%)	1.101 (44%)
(30,60]	2.084 (33%)	4.275 (67%)
(60,90]	1.165 (12%)	8.955 (88%)
(90, Inf]	732 (7%)	10.264 (93%)
Classificação da empresa (*NewExist*)	**Não**	**Sim**
Negócio Existente	3.787 (18%)	1.7643 (82%)
Novo Negócio	1.611 (19%)	6.952 (81%)
Zona de atuação (UrbanRural)	**Não**	**Sim**
Não declarado	808 (7%)	10.155 (93%)
Urbano	3.940 (25%)	11.649 (75%)
Rural	650 (19%)	2.791 (81%)
Programa de Empréstimo (*LowDoc*)	**Não**	**Sim**
Não	5.033 (19%)	21.008 (81%)
Sim	365 (9%)	3.587 (91%)
Ramo de Atuação (*NAICS*)	**Não**	**Sim**
Comércio	1.511 (22%)	5.113 (78%)
Indústria	929 (19%)	3.839 (81%)
Outros	1.146 (12%)	8.429 (88%)
Serviços	1.812 (20%)	7.214 (80%)

Fonte: Elaborado pelo autor

Percebe-se que independente do tempo de existência da empresa, a maior parte dos empréstimos concedidos foram adimplentes. Sobre o ramo de atuação no mercado, tem-se que as entidades das classes de comércio, indústria e serviços tiveram um percentual de inadimplentes próximos a 20%, já a classe que representa outros tipos de ramo obteve um percentual menor que 11%. A zona de atuação da empresa também foi um fator considerado na análise, nota-se na Tabela 3 que a maioria dos créditos concedidos foi destinado a empresas com atuação na zona urbana. Esta maioria apresentou uma

probabilidade maior de inadimplência (25%) do que as empresas da zona rural (18%) e até mesmo das empresas que não registraram a informação de localização (7%). Ademais, verifica-se que as organizações participantes do programa de empréstimo das entidades credoras tiveram um percentual de inadimplência menor (9%) em comparação com as corporações não participantes (19%).

O valor médio do crédito concedido pelos bancos ("***GrAppv***") é de US$ 87.809, com um desvio padrão no valor de US$ 282.942, indicando assim uma alta variação nessa variável, pois conforme apresentado na Tabela 4, o menor valor concedido foi de US$ 1.000 e no outro extremo da distribuição fortemente assimétrica, tem-se o maior crédito aprovado de US$ 5 milhões. Já o valor médio da cobertura de crédito aprovada pela SBA ("***SBA_Appv*"**) é US$ 60.487, também contendo uma alta variabilidade, pois o desvio padrão foi de US$ 228.407. Nota-se que as variáveis relacionadas a volumetrias de emprego ("***NoEmp", "CreateJob"*** e **"*RetainedJob"***) possuem uma alta assimetria e presença de *outliers*, pois em média o número de funcionários, a criação e retenção de empregos tendem a ser menor ou igual a 5 por empresa, contudo o desvio padrão para essas variáveis é maior ou igual a 90, indicando assim que a maior parte das organizações solicitantes são de pequeno porte, porém há registros de empresas de médio/grande porte na base.

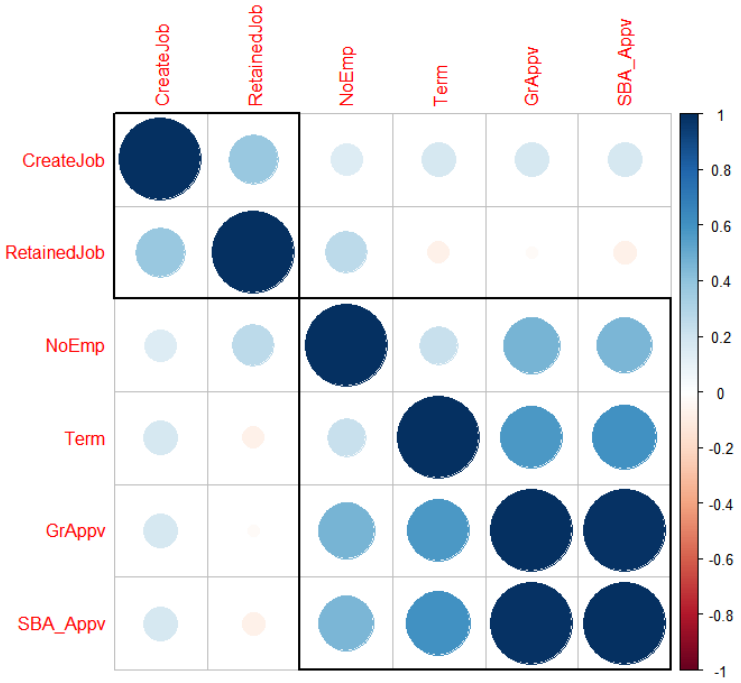
Tabela 4 - Medidas de resumo das variáveis quantitativas selecionadas para a modelagem para previsão da inadimplência

Variáveis	**Medidas de Resumo**				
	Mínimo	**Mediana**	**Média**	**Desvio Padrão**	**Máximo**
GrAppv	1.000	90.000	87.809	282.942	5.000.000
SBA_Appv	500	62.500	60.487	228.407	4.500.000
NoEmp	0	4	5	90	8018
CreateJob	0	0	1	209	8.800
RetainedJob	0	1	1	210	8.800

Fonte: Elaborado pelo autor

Ao analisar se as variáveis quantitativas apresentam um grau significativo de correlação, constata-se na Figura 14 que entre as variáveis ("***SBA_Appv*"**) e ("***GrAppv*"**) existe uma correlação positiva forte, o que de fato é esperado, pois quanto maior o valor do crédito aprovado pelo banco, maior também será o valor coberto do empréstimo pela instituição SBA. Deste modo, a fim de evitar o efeito de multicolinearidade, a variável ("***GrAppv*"**) foi retirada da etapa de ajuste dos modelos.

Figura 14 - Matriz de correlação das variáveis quantitativas selecionadas para a modelagem para previsão da inadimplência



Fonte: Elaborado pelo autor

5.3 Ajuste dos modelos e avaliação das performances

Os modelos (M1), (M2), (M3) e (M4) foram desenvolvidos respectivamente pelos algoritmos: regressão logística, *naive bayes*, *decision tree* e *random forest* apresentados no capítulo 4. No que se refere ao desempenho, esses modelos obtiveram uma acurácia acima de 80% na generalização, isto é, acertaram a classe de mais de 80% dos clientes inseridos no conjunto de teste. Também é importante destacar que a sensibilidade dos modelos foi superior a 0.95, indicando que em termos percentuais 95% dos clientes adimplentes foram identificados corretamente. Por outro lado, com relação a especificidade, a performance dos ajustes não foi tão boa assim, pois a maior taxa foi 0.64, implicando que em termos percentuais esses ajustes conseguiram classificar corretamente menos de 65% dos clientes inadimplentes do conjunto de teste.

Os modelos (M1) e (M2) tiveram uma performance semelhante entre si e inferior quando comparados com os demais, sobretudo na métrica da especificidade exposta na Tabela 5, acertando as classes de 84% dos clientes avaliados, em que 98% dos adimplentes foram classificados corretamente, e menos de 21% dos clientes inadimplentes foram previstos da maneira correta. Já o (M3) obteve uma acurácia de 89% dos registros, tendo a capacidade de acertar 96% dos adimplentes e 56% dos inadimplentes. O (M4) apresentou

o melhor desempenho entre os demais citados, pois sua acurácia foi igual 90%, classificando corretamente 96% dos clientes adimplentes e 64% dos clientes inadimplentes.

Tabela 5 - Métricas de avaliação da performance dos modelos na previsão da inadimplência (Cenário desbalanceado)

Modelo	Algoritmo	Acurácia	Sensibilidade	Especificidade	Precisão
M1	Logística	0,839	0,979	0,200	0,842
M2	*Naive bayes*	0,832	0,986	0,129	0,837
M3	*Decision tree*	0,894	0,966	0,562	0,910
M4	*Random forest*	0,904	0,963	0,640	0,924

Fonte: Elaborado pelo autor

O relato do bom desempenho na acurácia e o menor poder preditivo para uma determinada classe corrobora com a afirmação do Liu et al. (2019) sobre a ilusão de uma boa performance apresentada pelos algoritmos em cenários com dados desbalanceados, pois ao analisar o indicador de acurácia é possível supor que os modelos têm uma alta capacidade de discriminação, contudo averiguando a sensibilidade e especificidade, nota-se que eles não estão contendo poder preditivo para a classe de inadimplentes, o que torna inviável sua utilização, visto que na melhor situação proposta, cerca de 36% dos clientes inadimplentes foram classificados como adimplentes. Segundo Branco et al. (2016), a performance ilusória ocorre devido ao número maior de registros na classe de clientes adimplentes, logo os algoritmos tendem a aprender menos sobre os clientes inadimplentes, uma vez que sua participação no conjunto de dados é menor (18%).

Visando contornar o possível efeito do desbalanceamento dos dados, os modelos (M5), (M6), (M7) e (M8) foram desenvolvidos utilizando a técnica de *Undersampling e* os modelos (M9), (M10), (M11) e (M12) da técnica de *Oversampling*, seguindo respectivamente os algoritmos apresentados na Tabela 6. O auxílio das técnicas do balanceamento tende a propiciar que os algoritmos tenham um desempenho mais equilibrado. O treinamento em cenários balanceados ocasionou a redução da acurácia e sensibilidade, porém gerou um aumento considerável da especificidade de todos os ajustes, com exceção para o algoritmo *random forest*, que apenas teve redução no indicador de sensibilidade.

Tabela 6 - Métricas de avaliação da performance dos modelos na previsão da inadimplência (Cenários balanceados)

Modelo	Algoritmo	Amostragem	Acurácia	Sensibilidade	Especificidade	Precisão
M5	Logística	*Undersampling*	0,671	0,640	0,812	0,939
M6	*Naive bayes*		0,687	0,671	0,756	0,926
M7	*Decision tree*		0,845	0,836	0,885	0,971
M8	*Random forest*		0,905	0,901	0,925	0,982
M9	Logística	*Oversampling*	0,670	0,641	0,817	0,941
M10	*Naive bayes*		0,694	0,678	0,770	0,931
M11	*Decision tree*		0,824	0,801	0,932	0,982
M12	*Random forest*		0,947	0,950	0,941	0,985

Fonte: Elaborado pelo autor

Diferentemente dos demais ajustes, o algoritmo *Random forest* teve o indicador de acurácia melhorado utilizando as técnicas de *Undersampling*, e ainda mais utilizando *Oversampling*, segundo exposto na Tabela 7. Apresentando no cenário de redução da classe majoritária uma acurácia, sensibilidade e sensibilidade iguais a 90%, e no cenário de aumento da classe minoritária indicadores acima de 94%.

Tabela 7 - Métricas de avaliação da performance dos modelos advindos do *Random forest*

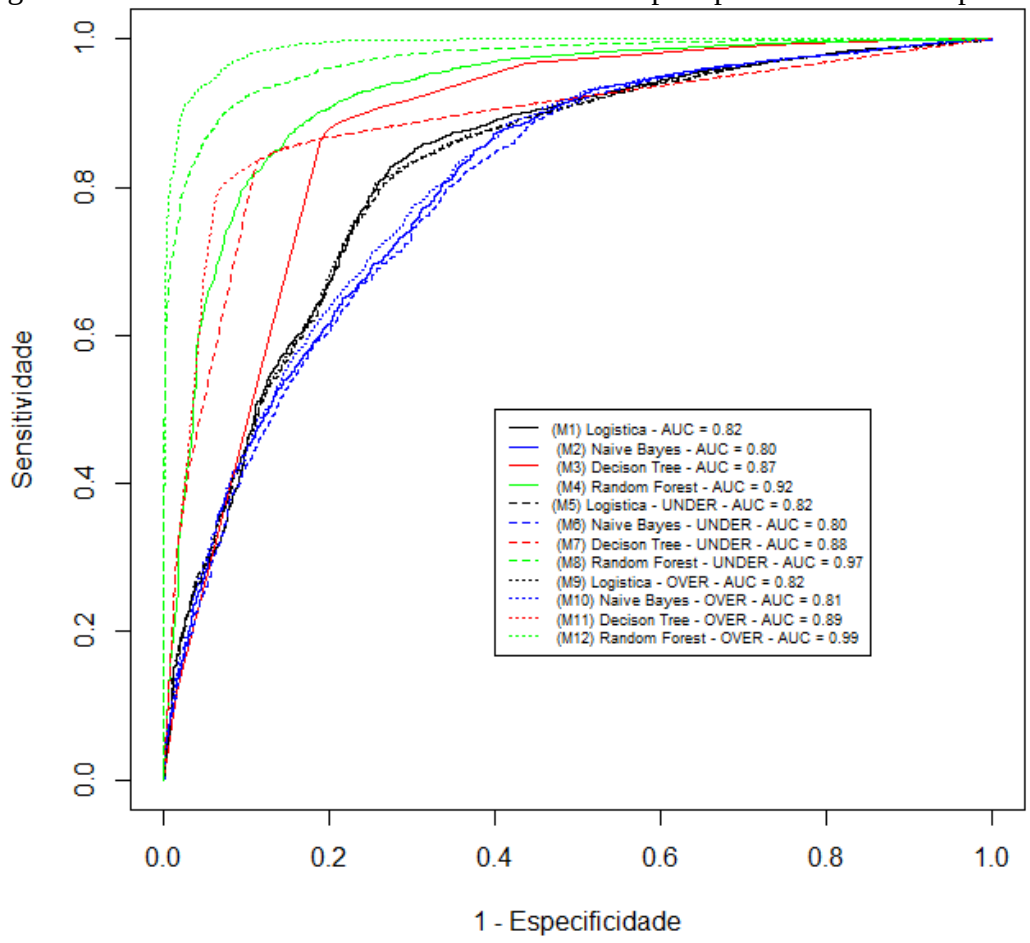
Modelo	Amostragem	Acurácia	Sensibilidade	Especificidade	Precisão
M4	-	0,904	0,963	0,640	0,924
M8	*Undersampling*	0,905	0,901	0,925	0,982
M12	*Oversampling*	0,947	0,950	0,941	0,985

Fonte: Elaborado pelo autor

Conforme apresentado na Figura 15, pela perspectiva da curva ROC observa-se que os algoritmos *Naive Bayes* e regressão logística apresentaram os piores desempenhos para esses dados, e a utilização das técnicas de balanceamento não ocasionou diferença significativa em suas performances, pois o indicador de AUC permaneceu em média 0.80 para os modelos de *Naive Bayes* e 0.82 para os da regressão logística. Por outro lado, o treinamento com dados balanceados provocou um ganho preditivo no algoritmo *decision tree.* Os modelos calibrados (M7) e (M11) constituíram, respectivamente, curvas com AUC igual a 0.88 e 0.89, enquanto o modelo treinado no cenário desbalanceado (M3) obteve um AUC de 0.87. Vale enfatizar que os treinamentos utilizando as técnicas de *undersampling* e *oversampling* ocasionaram uma redução máxima de 16% na sensibilidade dos modelos, em contrapartida aumentou a especificidade em até 70%, gerando assim uma melhor performance.

O algoritmo *random forest* foi responsável pelos ajustes com os melhores desempenhos, e do mesmo modo que no algoritmo *decision tree*, a aplicação de técnicas de balanceamento dos dados melhorou significativamente a sua performance. Todavia, nota-se na Figura 15 que para o *random forest*, o balanceamento ocasionou as menores reduções na sensibilidade. O modelo treinado no cenário desbalanceado (M4) obteve um AUC igual a 0.92, já o modelo treinado no cenário *undersampling* (M8) teve uma performance ainda melhor contendo um AUC igual a 0.97, e por fim, o modelo (M12) treinado no cenário *oversampling* foi o que apresentou o melhor desempenho entre todos os ajustes, com um AUC igual a 0.99.

Figura 15 - Curva ROC dos modelos desenvolvidos para previsão de inadimplência



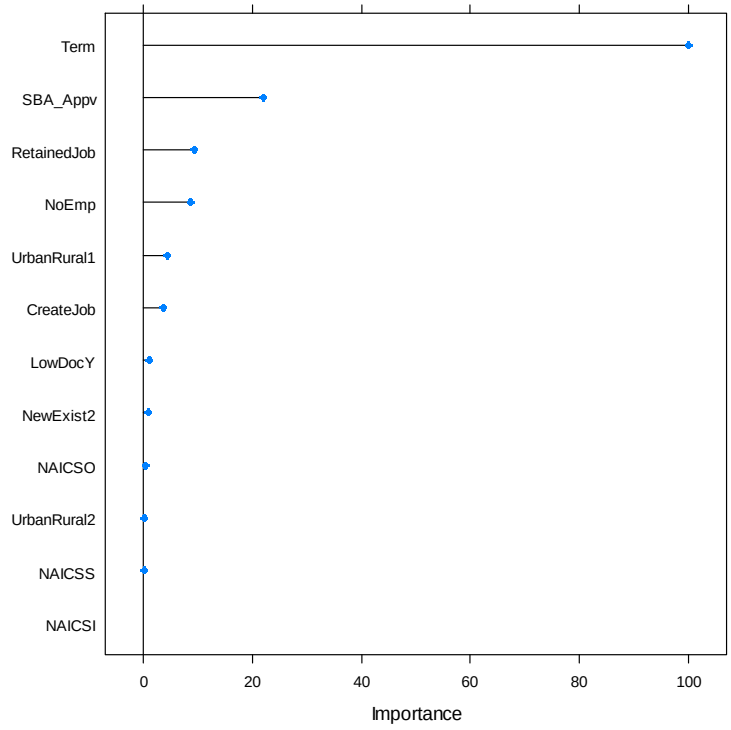
Fonte: Elaborado pelo autor

O baixo desempenho dos modelos obtidos por *Naive Bayes* reforçam os resultados apresentados por Santos (2013). Mesmo o indicador de AUC não sofrendo alteração significativa, nota-se que os modelos de *Naive bayes* e regressão logística treinados nos cenários balanceados tornaram-se mais eficientes na classificação de clientes

inadimplentes, estando em concordância com os resultados de Cordeiro (2020) e dos estudiosos Nguyen e Huynh (2020). O fato de que os melhores desempenhos foram observados utilizando *random forest* coincide com os estudos de Lakshmi e Kavilla (2018), Li et al (2020) e Sousa (2021). Ademais, dado que este algoritmo é classificado como *ensemble*, sua superioridade em comparação aos demais algoritmos individuais também é apontada pelos estudiosos Stelzer (2019), Cordeiro (2020), Nguyen e Huynh (2020).

O modelo (M12), considerado o melhor ajuste, foi desenvolvido pela união de 500 árvores de decisão que continham até 3 variáveis em suas ramificações e o parâmetro de aleatoriedade $mtry$ aplicado foi igual a 3.32. Entre as variáveis preditoras, verifica-se na Figura 16 que o atributo "***Term*"**, indicador do número de parcelas que o empréstimo foi dividido, teve grande importância para o modelo. O valor do empréstimo aprovado pela instituição SBA "***SBA_APPV***" também foi importante para o poder discriminatório. Além disso, averígua-se que as informações sobre o número de empregos retidos "***RetainedJob***", número de empregados "***Noemp***", criação de novos empregos "***CreateJob***" e a localização da empresa "***UrbanRural***" também contribuíram para classificação do status de *default* dos clientes.

Figura 16 - Nível de importância das variáveis para o modelo advindo do algoritmo *Random forest* e treinado no cenário *Oversampling* (M12)



Fonte: Elaborado pelo autor

6 CONSIDERAÇÕES FINAIS

No decorrer deste estudo buscou-se apresentar a aplicabilidade de algoritmos de *machine learning* para o desenvolvimento de modelos que identifiquem clientes propensos a não honrar com suas dívidas perante as entidades financeiras. Diante dos resultados apresentados, observa-se que, para os dados explorados, os modelos desenvolvidos no cenário desbalanceado passam a impressão de que as suas previsões acertam com frequência devido à qualidade dos ajustes, contudo, na verdade, é a predominância de uma das classes que faz com que os modelos acertem na maioria de suas classificações. Ou seja, os altos valores da acurácia em cenários desbalanceados não significam que o desempenho do modelo é satisfatório, para esta afirmação, faz-se necessário investigar também os indicadores de sensibilidade e especificidade. A utilização das técnicas de *Undersamplimg* e *Oversampling* proporcionou aos algoritmos uma etapa de treinamento mais equilibrada, resultando em uma melhor compreensão sobre os clientes inadimplentes. O treinamento em cenários balanceados ocasionou a redução da acurácia e sensibilidade na maior parte dos casos, porém gerou um aumento considerável da especificidade de todos os ajustes.

Pelo que foi exposto, observa-se que o algoritmo *Random Forest* obteve as melhores métricas de avaliação entre os algoritmos utilizados, independente do cenário de treinamento proposto. Ademais, utilizando a métrica de AUC, tem-se que este algoritmo treinado sobre a técnica de *Oversampling* resultou no melhor desempenho referente à generalização. Destaca-se que a presente pesquisa cumpriu com seus objetivos, e seus resultados podem contribuir para a comunidade acadêmica, uma vez que o tema proporciona a união do conhecimento teórico com a aplicação técnica em problemas reais do mercado de trabalho. Os resultados podem ser relevantes para gestores de modelagem de crédito em instituições financeiras, caso precisem direcionar o desenvolvimento de modelos ou traçar novos métodos de *credit score*.

Na perspectiva de trabalhos futuros, sugere-se a utilização de outros algoritmos de ML, sobretudo os atrelados ao método *ensemble.* Além disso, a exploração de outras técnicas para balanceamento dos dados também pode proporcionar discussões e resultados relevantes. Por fim, recomenda-se a utilização de diferentes bases de dados, a fim de averiguar a existência de padrão na aplicação das técnicas.

REFERÊNCIAS

ALVES, Marleide F.; LOTUFO, Anna Diva P.; LOPES, Mara Lúcia M. Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 1, n. 1, 2013.

ANICETO, Maísa Cardoso. **Estudo comparativo entre técnicas de aprendizado de máquina para estimação de risco de crédito**. 2016. Dissertação (Mestrado em Administração). Universidade de Brasília, Brasília.

BARELLA, Victor Hugo. **Técnicas para o problema de dados desbalanceados em classificação hierárquica**. 2016. Tese de Doutorado. Universidade de São Paulo.

BATISTA, Gustavo Enrique de Almeida Prado et al. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese de Doutorado. Universidade de São Paulo.

BOEHMKE, Brad; GREENWELL, Brandon. **Hands-on machine learning with R**. Chapman and Hall/CRC, 2019.

BORBA, Vítor Eduardo Galeão Borba de. **Aplicação de árvores de decisão na seleção de portfólios de ações**. 2021. Trabalho de conclusão de curso (Bacharelado em Estatística). Universidade Federal do Rio Grande do Sul.

BRANCO, Paula; TORGO, Luís; RIBEIRO, Rita P. A survey of predictive modeling on imbalanced domains. **ACM Computing Surveys (CSUR)**, v. 49, n. 2, p. 1-50, 2016.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

CORDEIRO, Tiago Vilas Boas. **Predição de default de empresas: técnicas de machine learning em dados desbalanceados**. 2020. Tese de Doutorado. Fundação Getúlio Vargas. São Paulo. 2020.

ANDRADE, Renato Hudson. **Avaliação de Risco de Crédito utilizando grupo de classificadores**. 2008. Dissertação (Mestrado em Estatística). Universidade Federal de Minas Gerais. Belo Horizonte. 2008.

FIGUEIRA, C. V. **Modelos De Regressão Logística**. 2006. Dissertação (Mestrado em Matemática). Universidade Federal do Rio Grande do Sul, [s. l.], 2006.

FONTELLES, Mauro José et al. Metodologia da pesquisa científica: diretrizes para a elaboração de um protocolo de pesquisa. **Revista paraense de medicina**, v. 23, n. 3, p. 1-8, 2009.

FORTI, Melissa. **Técnicas de machine learning aplicadas na recuperação de crédito do mercado brasileiro**. 2018. Tese de Doutorado. Fundação Getúlio Vargas. São Paulo. 2018.

GALLICCHIO, Cláudio et al. Abordagens Randomizadas de Aprendizado de Máquina: Desenvolvimentos e Desafios Recentes. Em: **ESANN**. 2017.

GISLASON, Pall Oskar; BENEDIKTSSON, Jon Atli; SVEINSSON, Johannes R. Random forests for land cover classification. **Pattern recognition letters**, v. 27, n. 4, p. 294-300, 2006.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. Random forests. In: **The elements of statistical learning**. Springer, New York, NY, 2009. p. 587-604.

KARCHER, Cristiane. **Redes Bayesianas aplicadas à análise do risco de crédito**. 2009. Tese de Doutorado. Universidade de São Paulo.

LAKSHMI, S. V. S. S.; KAVILLA, Selvani Deepthi. Machine learning for credit card fraud detection system. **International Journal of Applied Engineering Research**, v. 13, n. 24, p. 16819-16824, 2018.

LEMOS, Eliane Prezepiorski; STEINER, Maria Teresinha Arns; NIEVOLA, Julio César. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. **Revista de Administração-RAUSP**, v. 40, n. 3, p. 225-234, 2005.

LI, Jing-Ping et al. Machine learning and credit ratings prediction in the age of fourth industrial revolution. **Technological Forecasting and Social Change**, v. 161, p. 120309, 2020.

LI, Min; MICKEL, Amy; TAYLOR, Stanley. "Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines. **Journal of Statistics Education**, v. 26, n. 1, p. 55-66, 2018.

LIU, Shiyu et al. Early prediction of sepsis via smote upsampling and mutual information based downsampling. In: **2019 Computing in Cardiology (CinC)**. IEEE, 2019. p. Page 1-Page 4.

LOBATO, Tarcísio; CARVALHO, Brena. Proposta de um modelo ensemble para credit scoring. **Brazilian Journal of Development**, v. 7, n. 3, p. 24280-24297, 2021.

MARQUES, Roberto Ligeiro; DUTRA, I. N. Ê. S. Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. **Coppe Sistemas–Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil**, 2002.

MYLES, Anthony J. et al. An introduction to decision tree modeling. **Journal of Chemometrics: A Journal of the Chemometrics Society**, v. 18, n. 6, p. 275-285, 2004.

NEAPOLITAN, Richard E. Learning Bayesian networks. In: **Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining**. 2007. p. 1-1.

NGUYEN, Hung Ba; HUYNH, Van-Nam. On sampling techniques for corporate credit scoring. **Journal of Advanced Computational Intelligence and Intelligent Informatics**, v. 24, n. 1, p. 48-57, 2020.

OLIVEIRA, Diego Silveira Pacheco de. **Ensaios sobre a avaliação de risco de crédito soberano: determinantes, impacto sobre o desacordo das expectativas de câmbio e a capacidade de previsão de técnicas de Machine Learning**. 2020. Tese de doutorado. Universidade Federal Fluminense. Niterói. 2020.

PAIVA, Cláudia Costa Vieira. **Previsão da Inadimplência através da Regressão Logística**. 2015. Monografia (Especialização em Estatística). Universidade Federal de Minas Gerais. Belo Horizonte. 2015.

PEREIRA, Pedro Miguel Pinhal. **Análise de risco de crédito usando algoritmos de Machine Learning**. 2021. Dissertação (Mestrado em Matemática Financeira). Universidade de Lisboa. Lisboa. 2021.

PIEDADE, Márcio Palheta. **Uma abordagem de aprendizagem profunda que usa funções assimétricas para modelagem de pontuação de crédito no varejo**. 2020. Tese de doutorado. Universidade Federal do Amazonas. Manaus. 2020.

PINO, Francisco Alberto. A questão da não normalidade: Uma revisão. **Revista de economia agrícola**, v. 61, n. 2, p. 17-33, 2014.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Áustria. 2021.

SANTOS, André Luiz Abreu. **O classificador Naïve Bayes no contexto da análise de crédito**. 2013. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação). Universidade Federal do Maranhão. São Luís. 2013.

SANTOS, Hellen Geremias. et al. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. **Cadernos de Saúde Pública**, v. 35, 2019.

SOUSA, Túlio. **Estudo de Algoritmos de Machine Learning para Predição de Fraudes em Cartões de Crédito**. 2021. Trabalho de Conclusão de Curso. Pontifícia Universidade Católica de Goiás. Goiânia. 2021.

STELZER, Anna. Predicting credit default probabilities using machine learning techniques in the face of unequal class distributions. **arXiv preprint arXiv:1907.12996**, 2019.

WEBB, Geoffrey I.; KEOGH, Eamonn; MIIKKULAINEN, Risto. Naïve Bayes. **Encyclopedia of machine learning**, v. 15, p. 713-714, 2010.

YADAV, Sanjay; SHUKLA, Sanyam. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: 2016 IEEE 6th International conference on advanced computing (IACC). IEEE, 2016. p. 78-83.