UNIVERSIDADE DE BRASÍLIA
DEPARTAMENTO DE ADMINISTRAÇÃO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE (FACE).

FERNANDA SANTOS AMORIM

Previsão de Indícios de Fraude em Fundos de Pensão utilizando Modelos de Aprendizado de Máquina Supervisionados e Técnicas de Balanceamento de Dados

Brasília, DF
2021

FERNANDA SANTOS AMORIM

Previsão de Indícios de Fraude em Fundos de Pensão utilizando Modelos de Aprendizado de Máquina Supervisionados e Técnicas de Balanceamento de Dados

Dissertação apresentada ao Programa de Pós-Graduação em Administração da Universidade de Brasília para obtenção do Título de Mestre em Administração no Eixo Temático de Finanças e Métodos Quatitativos.

Brasília, DF
2021

FERNANDA SANTOS AMORIM

Previsão de Indícios de Fraude em Fundos de Pensão utilizando Modelos de Aprendizado de Máquina Supervisionados e Técnicas de Balanceamento de Dados

Dissertação apresentada ao Programa de Pós-Graduação em Administração da Universidade de Brasília para obtenção do Título de Mestre em Administração no Eixo Temático de Finanças e Métodos Quatitativos.

Área de concentração:
Finanças e Métodos Quantitativos

Supervisor:
Prof. Dr. Pedro Henrique Melo Albuquerque

Brasília, DF
2021

FICHA

CATALOGRÁFICA

Amorim, Santos Fernanda
Previsão de Indícios de Fraude em Fundos de Pensão utilizando Modelos de Aprendizado de Máquina Supervisionados e Técnicas de Balanceamento de Dados / . – Brasília, DF, 2021. 94 p.

Dissertação (Mestrado) — Universidade de Brasília - Departamento de Administração.

1. Fraudes. 2. Dados Desbalanceados. 3. Aprendizado de Máquina. I. Amorim, Santos Fernanda II. Universidade de Brasília.

AGRADECIMENTOS

Agradeço a Deus, a quem devo toda honra e toda glória por cada detalhe que ocorre na minha minha vida.

Ao meu orientador, Pedro Albuquerque, que não só viu potencial em mim desde o dia que eu entrei na sala pela primeira vez, como transformou este potencial em aprendizado. Devo muito da minha vida acadêmica e profissional a você, Pedro.

À minha família, Mônica, Daniel, Daniel Júnior e Arthur, obrigada pelo suporte, por aguentar o meu estresse durante esses anos de estudo e por não me deixarem desistir em momento nenhum. Vocês são a minha melhor parte.

Aos meus amigos que foram apoio e coragem em tempos de dificuldade, que foram ouvidos em tempos de angústia e me deram forças para continuar mesmo quando o cansaço já estava no ápice. Não tenho como citar todos, mas queria deixar alguns nomes em especial, Ana Júlia Ribeiro, Monique Renault, Eloise Mary, Gabriela Freire, Marcus Barbosa, Josué Cardoso, Rafael Xavier, Ana Carolina Tigre, Jhon Heider, Ana Luiza Egito, Ana Júlia Oliveira, Luísa Alfenas, Gustavo Monteiro, Gabriela Nobre, Manuela Melo e Bruno Ferreira.

Aos meus companheiros de Universidade de Brasília e membros do Laboratório de Aprendizado de Máquina em Finanças e Organizações. Aprendo muito com vocês, muito obrigada.

Aos amigos do Summer Acadêmico do Itaú Unibanco, também conhecidos como melhores colegas de trabalho que eu poderia pedir. Obrigada por me ouvirem. Dividir a caminhada acadêmica com vocês fez desse caminho muito mais leve.

RESUMO

Fraudes Financeiras têm se tornado alvo de diversos estudos, devido aos impactos (principalmente econômicos), que estas podem causar a empresas, países e pessoas físicas. Detecção de Fraudes é a área de estudos que procura identificar atividades fraudulentas. Estas análises são feitas dentro de conjuntos de dados que são muito desbalanceados devido à baixa ocorrência dos acontecimentos alvo, isto é, existem classes de dados que ocorrem com maior frequência (classes majoritárias) que outras (classes minoritárias). Os dados que ocorrem com menor frequência são conhecidos como eventos raros e podem ser observados em diversas áreas de estudo como medicina (doenças raras), sistemas de rede (detecção de intrusos), meteorologia (desastres naturais), finanças (fraudes, falência). O estudo proposto tem como objetivo avaliar o desempenho de Modelos Supervisionados de Aprendizado de Máquina para dados desbalanceados de Indício de Fraudes em Fundos de Pensão utilizado Técnicas de Balanceamento de dados. Os dados utilizados foram cedidos pela Superintendência Nacional de Previdência Complementar (PREVIC). Para Seleção de Variáveis, foi usado Análise de Componentes Principais. Os modelos utilizados foram: Regressão Logística, *Random Forest*, Máquina de Suporte Vetorial e Redes Neurais. As Técnicas de Balanceamento utilizadas foram *Random Undersampling*, SMOTE e SMOTETomek. Com os testes realizados, este estudo recomenda a utilização do *Random Forest* como Modelo de Aprendizado de Máquina, ajustando o desbalanceamento da base com o SMOTE, por ter apresentado os melhores resultados de acordo com as Métricas de Avaliação utilizadas.

ABSTRACT

Financial frauds has become the target of several studies, due to its impacts (which are mainly economical) that can cause to companies, countries and individuals. Fraud Detection is the field of study that seeks to identify fraudulent activities. These analyzes are made within datasets that are very unbalanced due to the low occurrence of the target events, that is, there are data classes that occur more frequently (majority classes) than others (minority classes). The data that occur less frequently are known as rare events and can be seen in several fields of studies such as medicine (rare diseases), network systems (intrusion detection), meteorology (natural disasters), and finance (fraud, bankruptcy). The proposed study aims to evaluate the performance of Supervised Machine Learning Models for unbalanced data of Fraud Indication in Pension Funds using Data Balancing Techniques. The data used were provided by the National Superintendency of Complementary Pension (PREVIC). For Variable Selection, Principal Component Analysis was used. The models used were: Logistic Regression, Random Forest, Vector Support Machine and Neural Networks. The Balancing Techniques used were Random Undersampling, SMOTE and SMOTETomek. With the tests performed, this study recommends the use of *Random Forest* as a Machine Learning Model, adjusting the base unbalance with SMOTE, as it presented the best results according to the Evaluation Metrics used.

LISTA DE FIGURAS

3.1 Mapa de Calor - PCA 42

3.2 Separador de Máxima Margem Fonte: Soman, Loganathan e Ajay (2009) 46

4.1 Métricas - Regressão Logística 52

4.2 Métricas - *Random Forest* 53

4.3 Métricas - Máquina de Suporte Vetorial 54

4.4 Gráfico de Comportamento do *Geometric Mean Score* Acumulado para Diferentes tipos de Validação 58

4.5 Gráfico de Comportamento do *Geometric Mean Score* Absoluto para Diferentes tipos de Validação 59

4.6 Resultado *Geometric Mean Score* 63

4.7 Resultado Teste de Stress - *Random Undersampling* 67

4.8 Resultado Teste de Stress - SMOTE 68

4.9 Resultado Teste de Stress - SMOTETomek 69

LISTA DE TABELAS

2.1 Classificação de Fraude Fonte: Adaptado de Ngai et al. (2011) 18

3.1 Interpretabilidade dos Componentes Principais 42

4.1 Matriz de Confusão 51

4.2 Métricas - Regressão Logística 52

4.3 Métricas - Máquina de Suporte Vetorial 54

4.4 Média do *Geometric Mean Score* 63

4.5 Variância do *Geometric Mean Score* 64

6.1 Dicionário de Variáveis 80

6.2 Valores do Beta da Regressão LASSO 83

6.3 Valores do Beta da Regressão Ridge 85

6.4 Combinação dos Modelos x Siglas 86

SUMÁRIO

1 Introdução **11**

1.1 Problema da Pesquisa 12

1.2 Motivação/Justificativa 15

2 Referencial Teórico **17**

2.1 Fraudes 17

2.1.1 Fraudes em Sistemas Financeiros 17

2.1.2 Aprendizado de Máquinas em Fraudes 20

2.2 Modelos de Aprendizado de Máquina para Eventos Raros 23

2.2.1 Modelos Supervisionados para Eventos Raros 26

3 Metodologia **34**

3.1 Análise Qualitativa 35

3.2 Base de Dados 35

3.3 Seleção de Variáveis 38

3.3.1 Regressão LASSO e Regressão Ridge 38

3.3.2 Análise de Componentes Principais 39

3.4 Modelos de Classificação Supervisionados 43

3.4.1 Regressão Logística 44

3.4.2 Random Forest 44

3.4.3 Máquina de Suporte Vetorial 45

3.4.4 LightGBM 47

3.4.5 Redes Neurais 48

4 Resultados **50**

4.1 Simulação 50

4.2 Técnicas de Balanceamento 55

4.2.1 *Random Undersampling* 59

4.2.2 SMOTE - *Synthetic Minority Over-sampling Tchnique* 60

4.2.3 SMOTE Tomek Links 60

4.3 Previsão dos Modelos Supervisionados 61

4.4 Combinação de Modelos 65

5 Conclusão **71**

6 Apêndice **76**

Referências Bibliográficas **87**

1 INTRODUÇÃO

O estudo de fraudes é importante e tem se tornado uma questão frequente em pesquisas na área de finanças, tendo em vista os possíveis prejuízos que podem ser causados às empresas (KOU et al., 2004). As fraudes compreendem parcelas muito pequenas de informações se considerarmos o conjunto completo de dados, devido a sua baixa ocorrência em estados normais, isto é, as classes que descrevem fraudes estão em muito menor quantidade que outras classes de dados. Na literatura, informações que possuem baixa incidência tem várias nomenclaturas e, consequentemente significados diferentes tais como evento raro, *outlier* ou anomalia, e podem surgir em diferentes áreas como, defeitos em software, intrusos em sistemas de rede, desastres naturais, células cancerígenas e fraudes financeiras (HAIXIANG et al., 2017; BEYAN; FISHER, 2015).

Diversos tipos de técnicas são utilizadas na detecção e/ou na previsão de fraudes e desses tipos de eventos, como *data mining*, modelos estatísticos e de inteligência artificial. Detectar e evitar estes eventos raros em finanças auxiliam as empresas a prevenir perdas e possibilita que organizações não percam credibilidade no mercado por envolvimento em escândalos financeiros, por exemplo. Contudo o avanço tecnológico propicia uma maior sofisticação dos ataques fraudulentos em finanças, por exemplo em fraudes com o uso de criptomoedas que dificilmente são rastreadas. Pensando na construção de novos padrões de ataques fraudulentos, faz-se necessário o uso de modelos mais aprimorados na detecção ou previsão, principalmente porque distribuições normais já assimilam pouca massa de probabilidade em valores extremos, tendo em vista que estes valores se encontram nas caudas das distribuições, e conforme Sharpe (1970) modelos usuais de previsão acabam desconsiderando estas probabilidades por serem muito baixas.

Este estudo propõe a aplicação de Modelos de Aprendizado de Máquina (AM)[1], para previsão de Fraudes em Fundos de Pensão. Fundos de Pensão são fundos ad-

[1] do inglês, *Machine Learning*

ministrados por empresas ou associações que tem finalidade de compor a previdência complementar da aposentadoria dos funcionários da empresa. Estudos relacionando Fraude e Aprendizado de Máquina já são presentes na literatura, principalmente com aplicações em fraudes de cartão de crédito (BHATTACHARYYA et al., 2011). Entretanto, poucos são estudos relacionados à fraude em fundos de pensão.

Além da importância de detecção deste tipo de fraude para evitar perdas financeiras de empresas e pessoas físicas, existe o desafio de previsão de eventos de baixa probabilidade de ocorrência para as técnicas de Aprendizado de Máquina. Este tipo de modelagem, em sua maioria, busca por generalização das classes majoritárias dos conjuntos de dados e isso gera uma dificuldade na modelagem das classes minoritárias, isto é, dos eventos raros (BEYAN; FISHER, 2015).

Ademais, Fundos de Pensão é o termo comum para Entidades Fechadas de Previdência Complementar (EFPC) que são fundações operadas por empresas e associações que têm como finalidade a administração de planos de benefícios, criados por empresas, para garantir aos seus empregados um complemento na aposentadoria oferecida pelo Regime Geral de Previdência Social do Instituto Nacional de Seguridade Social (INSS). As EFPCs são mantidas por meio de contribuições dos empregadores e dos empregados e a manutenção é feita pelos planos de benefícios, isto é, os planos que definem como o valor retorna para o participante. No caso de EFPCs que são geridas por associações, ocorre da mesma forma, porém os beneficiários são os associados [2].

As perdas causadas por fraudes em fundo de pensão atingem diretamente os contribuintes e, por isso, estudos em previsão destas fraudes são benéficos para evitar prejuízos para pessoas físicas e para empresas. Sob esta lógica, aprimorar as metodologias de detecção com o uso de Modelos de Aprendizado de Máquina é uma justificativa prática para o presente trabalho, tendo em vista que um dos objetivos deste estudo é a aplicação de Modelos Supervisionados para prever Indícios de Fraude em Fundos de Pensão.

1.1 Problema da Pesquisa

Para detecção de fraudes em fundos de pensão, é preciso lidar com a presença de eventos raros em bases de dados e, por consequência, problemas computacio-

[2] http://www.previc.gov.br/a-previdencia-complementar-fechada/sobre-o-setor

nais e estatísticos devido a bases de dados desbalanceadas. Dados desbalanceados ocorrem quando um conjunto de dados apresenta uma ou mais classes com um número muito maior de exemplos que outras, por exemplo, na base de dados usada no estudo de Raza e Qayyum (2019) para detecção de anomalia, os dados normais representaram 99,83% enquanto 0,17% representavam os dados que o estudo se propunha a detectar. Modelos tradicionais de classificação não são recomendados para dados desbalanceados (KING; ZENG, 2001), porque os resultados da classificação não vão condizer com a realidade dos dados, o modelo considera as classes majoritárias e os resultados para as classes minoritárias apresentam resultados enviesados minoritárias apresentam resultados enviesados (PHUA et al., 2010). Isso acontece porque o processo de aprendizado leva em consideração as classes majoritárias dos dados para generalização do aprendizado e, assim, produz uma previsão que consiga explicar grande parte dos dados (HAIXIANG et al., 2017), mas não necessariamente os eventos raros. Eventos raros também podem ser tratados como ruído, isto é, dados que são inconsistentes ao conjunto completo de dados que pertencem. Isso ocorre devido ao processo de aprendizagem do algoritmo, ruídos tem padrão anômalo como eventos raros, logo o algoritmo identifica erroneamente os dados nas classes minoritárias como ruído (HAIXIANG et al., 2017; BEYAN; FISHER, 2015). Dessa forma, o campo de estudo de eventos raros por vezes compartilha técnicas do campo de detecção de *outliers* uma vez que o objetivo final é reconhecer observações que não pertencem a "normalidade" dos dados.

Para solucionar as dificuldades envolvidas, (BEYAN; FISHER, 2015) cita algumas abordagens que podem ser utilizadas como modificações no próprio algoritmo para auxiliar o modelo de classificação a convergir e gerar resultados consistentes com o nível de desbalanceamento, geralmente estipulando pesos diferentes entre as classes. Outra possibilidade é o uso de abordagens sensíveis a custo [3], atribuindo custos diferentes para os dados de treinamento e suas classes. Também é possível fazer reamostragem dos dados de forma a balancear as classes seja com *oversampling*, isto é, aumento das classes minoritárias imputando novos dados, ou com *undersampling* que consiste na diminuição dos dados da classes majoritárias, retirando os dados da análise. Esta técnica funciona como pré-processamento dos dados e possui como dificuldade de aplicação a escolha de qual forma de amostragem é mais adequada para o tipo de estudo. Outras técnicas que também são utilizadas para solucionar as questões na aplicação de modelos em eventos raros são os Classificadores *Ensemble*,

[3] do inglês, *costs-sensitive*

como o *Bagging* e o *Boosting*.

Com intuito de contextualizar a importância desse campo de estudo em finanças com um exemplo real, em 2019, uma fraude em quatro fundos estatais deflagrada pelo Ministério Público e Polícia Federal gerou perdas de R$ 8 bilhões e mais de 630 mil trabalhadores foram prejudicados, isso mostra que há uma alta magnitude dos danos financeiros que podem ser causados por Fraudes em Fundos de Pensão fazendo com que a detecção destes eventos seja interessante para evitar prejuízos e conservar a credibilidade de empresas públicas e privadas [4]. Este estudo pretende fazer uma aplicação de Modelos de AM e de Técnicas de Balanceamento para prever fraudes em Fundos de Pensão. Eventos deste tipo ocorrem com baixa frequência em grandes conjuntos de dados e, por isso, são considerados eventos raros na área de Finanças. O problema de pesquisa deste estudo consiste em analisar o desempenho de modelos de Aprendizado de Máquina Supervisionados para previsão de fraudes e propor técnicas que possam auxiliar na previsão deste tipo de evento.

Sob ótica dos Modelos de Aprendizado de Máquina, podemos considerar que os modelos supervisionados se enquadram em abordagem reativa em relação aos dados, isto é, a Abordagem Reativa é aquela em que os dados são coletados após a ocorrência do evento e devido a isso, os dados são então rotulados. Estes dados rotulados de maneira reativa são resultado de ações gerenciais *a posteriori*, isto é, espera-se o evento ocorrer para tomar uma decisão relacionada ao futuro. A Abordagem Proativa no entanto antecipa a análise utilizando dados não rotulados, geralmente há uma hipótese econômica dos acontecimentos desses eventos e utilizam os dados para auxiliar na tomada de decisão futura (SARKAR et al., 2020). Este estudo focou em Abordagens Reativas em relação aos dados, isto é, o foco da metodologia de análise será em Modelos Supervisionados. Neste cenário, este estudo se propõe a responder a seguinte pergunta de pesquisa: “Sob ótica das Abordagens Reativa e do desbalanceamento das bases de dados de eventos raros, quais Modelos de Aprendizado de Máquina e Técnicas de balanceamento tem melhores resultados para Previsão de Fraude em Fundos de Pensão?”

[4] https://veja.abril.com.br/politica/greenfield-pede-r-4-bilhoes-por-fraudes-em-fundos-de-pensao/

1.2 Motivação/Justificativa

Fundos de Pensão são um tipo de investimento feito por empresas privadas ou públicas também chamado de Entidades Fechadas de Previdência Complementar (EFPC) e são reguladas pela Lei Complementar Nº 109/2001 que diz em seu Art 2o "O regime de previdência complementar é operado por entidades de previdência complementar que têm por objetivo principal instituir e executar planos de benefícios de caráter previdenciário, na forma desta Lei Complementar". Nos últimos anos, muitas operações da Polícia Federal e do Ministério Público Brasileiro deflagraram uma série de atividades fraudulentas de empresas trazendo a tona milhões de reais que foram usados de propina (Agência Brasil, 2018).

No Brasil, este tipo de Fundo é supervisionado pela Superintendência Nacional de Previdência Complementar (PREVIC) que trabalha em conjunto com o Ministério Público e a Polícia Federal e tem por missão "atuar na supervisão dos fundos de pensão de forma ágil, eficiente e transparente, com o objetivo de assegurar higidez e confiabilidade ao sistema de previdência complementar fechada" . Previdência complementar fechada faz parte do sistema de previdência social brasileiro como uma dar formas de poupança de longo prazo de funcionários, além disso, é maneira de diversificar a capacidade de investimentos do país, criando novas fontes de financiamento. A Previc analisa o comportamento dos Fundos de Pensão afim de encontrar indícios da fraude para indicar que seja feita uma investigação mais profunda por meio do Ministério Público e Polícia Federal (PREVIC, 2020). Como justificativa prática para este estudo, é possível observar que as técnicas utilizadas para investigar os fatos são manuais e com análise humana, logo a aplicação de Modelos de Aprendizado de Máquina pode ser uma alternativa para modernização do processo de investigação trazendo novos *insights* e auxiliando os Auditores na Tomada de Decisão.

O estudo de técnicas para previsão de fraudes em geral são importantes para o futuro das empresas e observando o cenário brasileiro de fundos de pensão. Devido às atuais intervenções das operações para deflagrar atos fraudulentos, um modelo que consiga prever possíveis fraudes nesta área pode prevenir ações de corrupção, lavagem de dinheiro e possíveis prejuízos às empresas envolvidas, demonstrando que esta pesquisa possui justificativa prática de aplicação, principalmente no contexto brasileiro. O estudo de Padula e Albuquerque (2018) mostra a importância de pesquisa relacionadas a prevenção de corrupção no contexto brasileiro realizando Estudo de Eventos para verificar o efeito dos acontecimentos decorrentes da Operação Lava-

Jato em quatro ativos de empresas estatais que compõe o Índice BM&FBOVESPA. É possível observar a influência negativa que a corrupção causa no mercado financeiro com a queda nos valores das ações nos retornos anormais e conseguinte desvalorização no curto prazo da Bolsa de Valores Brasileira.

Modelos de Aprendizado de Máquina são utilizados em diversas aplicações para detecção de fraude como em fraude de cartão de crédito (BHATTACHARYYA et al., 2011; FU et al., 2016), fraude bancária (ABDELHAMID; KHAOULA; ATIKA, 2014), fraude de seguros (KIRLIDOG; ASUK, 2012). Entretanto, para fraude em fundos de pensão não são encontrados muitos estudos com aplicação de Aprendizado de Máquina, mostrando a importância do estudo para a área como justificativa teórica de contribuição científica. Além disso, como um dos objetivos deste estudo é lidar com o desbalanceamento da bases de dados, outra justificativa teórica importante para os estudos da área é apresentar como o desbalanceamento afeta os Modelos de Aprendizado de Máquina e como é possível contornar o problema de classificação com Pré-Processamento e Técnicas de Balanceamento das bases de dados.

2 REFERENCIAL TEÓRICO

2.1 Fraudes

2.1.1 Fraudes em Sistemas Financeiros

O avanço da tecnologia que nos proporciona muitas facilidades de comunicação e poder de compra também possibilita a criação de novas possibilidades de se burlar os sistemas vigentes, ou seja, novas formas de criminosos praticarem fraudes (BOLTON; HAND, 2002). Uma das principais aplicações observadas na literatura de Eventos Raros dentro de Finanças é a identificação de fraude em sistemas financeiros e suas consequências, como por exemplo, fraudes como: de cartão crédito, bancária, agência de seguro e entre outros tipos. Fraude consiste em um ato contrário a lei, as regras ou as políticas de uma organização com a intenção de se obter benefício financeiro (WANG et al., 2006; PHUA et al., 2010). Para Ngai et al. (2011) que considerou a classificação de Fraude Financeira do FBI (*Federal Bureau of Investigation*), Fraudes Financeiras podem ser classificadas em dois níveis principais: alto e baixo. Fraudes de nível alto são fraudes financeiras propriamente ditas, como fraudes bancárias, fraudes de seguros, fraudes de valores mobiliários e outras fraudes financeiras relacionadas. Fraudes de nível baixo compreendem atividades fraudulentas, isto é, fraude hipotecária, confisco de bens/lavagem de dinheiro, fraude em assistência médica, fraude corporativa, fraude em *marketing* em massa e fraude previdenciária. Com esta classificação, é possível separar fraudes financeiras em quatro grupos principais e associar as atividades fraudulentas a cada um destes grupos como pode ser visto no Quadro 2.1:

Fraude Financeiras	Atividade Fraudulentas
Fraude Bancária	Fraude Hipotecária, Confisco de Bens, Lavagem de Dinheiro
Fraude de Seguros	Fraude em Assistência Média (Plano de Saúde), Fraude de Seguros de Automóvel
Fraudes de Títulos e Mercadorias	Operações com títulos
Fraude por Gestão Temerária	Fraudes em Fundos de Pensão
Outras Fraudes Financeiras	Fraude fiscal, Fraude Corporativa, Fraude de Marketing de Massas

Tabela 2.1 – Classificação de Fraude
Fonte: Adaptado de Ngai et al. (2011)

Detecção de Fraudes é uma área que estuda técnicas que buscam a identificação de comportamentos fraudulentos dentro de um conjunto de dados. Estes estudos ajudam a diminuir trabalhos manuais de triagem e verificação, automatizando processos, e tem se tornado uma aplicação muito visada para órgãos governamentais e indústrias de grande porte (PHUA et al., 2010). Existe uma série de aplicações comuns dentro dos estudos de detecção de fraude, como fraude de cartão de crédito, que são identificadas principalmente pela verificação dos registros de transações baseado no comportamento do cliente; fraude de telefone celular, principalmente relacionado com serviços de conta telefônica. Para detectar este tipo de fraude são feitos estudos de desvio de perfil das contas dos clientes, caso as contas apresentem valores que diferem muito do perfil de um usuário normal; fraude de seguros, que movimentam grande quantidade de recursos, esse tipo de fraude é extremamente difícil de ser detectada e geralmente são usados mecanismos de análise de documentos suspeitos para identificar as fraudes; e fraudes de *insider trading*, que ocorrem quando um *trader* utiliza alguma informação privilegiada para obter lucro na hora de realizar uma transação na bolsa de valores (AHMED; MAHMOOD; ISLAM, 2016; HAIXIANG et al., 2017; ALBUQUERQUE et al., 2019). Há também as fraudes previdenciárias que estão mais relacionadas com fraudes fiscais, um exemplo desta última consiste na tentativa de sonegação de algum imposto devido. Nos últimos 16 anos no Brasil, R$ 5,5 bilhões foram fraudados do Instituto Nacional do Seguro Social (INSS), mostrando que a fraude previdenciária lesou muito os cofres públicos (Correio Braziliense, 2019). As Fraudes relacionadas a Fundos de Pensão são classificadas como Fraudes por Gestão Temerária, isto é, fraudes praticadas por uma gestão indevida de instituições financeiros, expondo-as a riscos e possíveis prejuízos.

Detecção de Fraude é de extrema importância para a continuidade das empresas, tendo em vista que as consequências financeiras envolvidas em uma fraude podem ser devastadoras não só para uma organizações, como para a sociedade econômica e o governo (NGAI et al., 2011; BARMAN et al., 2016). Fraudes desestabilizam economias e

reduzem a confiança na indústria. Em 2008, Bernard Madoff que, na época, era presidente da empresa de investimentos Ponzi foi protagonista de uma das maiores fraudes financeiras dos Estados Unidos no valor de R$ 65 bilhões de dólares em fraudes bancárias, fiscais e lavagem de dinheiro, gerando prejuízos para diversas empresas Norte-Americanas (BARMAN et al., 2016; BHATTACHARYYA et al., 2011). Uma reportagem da BBC de 2007 mostrou que as perdas causadas por fraude em seguros no ano mencionado no Reino Unido chegou a 1,6 bilhões de libras (BBC News, 2007). Esses dois casos citados mostram a magnitude do que fraudes causam não só nas empresas, como também na economia dos países afetados.

Existem abordagens humanas feitas por auditoria que até são eficazes para detecção de fraude, entretanto não são tão eficientes e confiáveis pela dificuldade envolvida na detecção destes eventos. Atualmente, abordagens com uso de mineração de dados (*Data Mining*) e de *Machine Learning* tem sido amplamente utilizado na área para solucionar os problemas de detecção de eventos raros (RAZA; QAYYUM, 2019; NIU; WANG; YANG, 2019; WEST; BHATTACHARYA, 2016).

A grande dificuldade envolvida na detecção de eventos utilizando *Data Mining* ou *Machine Learning* encontra-se no desbalanceamento das bases de dados. Isso se dá porque fraudes são eventos raros, ou seja, têm uma incidência muito baixa de ocorrência em grandes volumes de dados, o que dificulta a interpretação e aprendizado, principalmente em modelos que tem como base a generalização. Este problema ocorre principalmente em métodos que utilizam abordagem supervisionada, tais métodos consistem em prever fraudes já previamente detectadas, isto é, utilizam uma abordagem reativa de generalizar o padrão de comportamento das fraudes para detectar novos eventos que possuem comportamento semelhante, neste caso, as bases de dados precisam ter uma classificação prévia para serem treinados. Um exemplo de modelos supervisionados são os modelos de Árvore de Decisão[1] ou os modelos Máquinas de Suporte Vetorial[2] com uso de Classificadores *Ensemble* como *Bagging e Boosting* (RAZA; QAYYUM, 2019; CODY; FORD; SIRAJ, 2015; PATEL; GOND, 2014; GALAR et al., 2011).

Além dos métodos supervisionados, também são aplicados para detecção e previsão de fraudes métodos não supervisionados que tem por objetivo identificar as fraudes antes que estas aconteçam, neste caso os dados não são previamente rotulados e o algoritmo precisa encontrar a estrutura dos dados para detectar os dados que

[1] do inglês, *Decision Trees*
[2] do inglês, *Support Vector Machine*

são as anomalias (*outliers*) em relação ao conjunto total (RAZA; QAYYUM, 2019; CHANDOLA; BANERJEE; KUMAR, 2009; NIU; WANG; YANG, 2019). A dificuldade envolvida nas aplicações de métodos supervisionados ocorre quando os dados anômalos apresentam características muitos semelhantes aos dados normais e, por isso, são muito difíceis de discriminar a diferença entre os dados anômalos e os dados normais (HAIXIANG et al., 2017).

2.1.2 Aprendizado de Máquinas em Fraudes

Modelos de Aprendizado de Máquina têm sido usados amplamente em Finanças, em áreas como Previsão de Falência (BARBOZA; KIMURA; ALTMAN, 2017), Precificação de Opções (CULKIN; DAS, 2017), Previsão de valores no mercado financeiro (HENRIQUE; SOBREIRO; KIMURA, 2019), Detecção e Previsão de Carteis (SILVEIRA et al., 2021) e na área de Fraudes Financeiras não é diferente. Dentro dos vários tipos de fraudes tem sido realizados estudos com aplicação de métodos de Aprendizado de Máquina com objetivo de melhorar a detecção destes eventos (PEROLS, 2011). Alguns exemplos de estudos utilizando Aprendizado de Máquina em Fraudes serão apresentados nesta Seção. A maioria das aplicações é em Fraudes de Cartão de Crédito e existem poucos estudos na área voltados para Fraude em Fundos de Pensão que é o foco do presente estudo.

O uso de cartão de crédito tem se tornado comum nos últimos anos e, com o crescente uso em transações comerciais, também há um crescente no número de fraudes financeiras em transações de cartão de crédito. Na tentativa de resolver esse problema, Fu et al. (2016) utiliza Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNN) para detectar fraudes de cartão de crédito. Uma das maiores dificuldades da aplicação do modelo é o desbalanceamento das bases de dados para fraude. Para resolver essa situação, o autor simulou dados (raros) fraudulentos a partir dos dados reais com uma metodologia de amostragem baseada em custos, dessa forma foi possível comparar o número de fraudes com o número de transações legítimas para treinar a CNN a detectar os eventos. O resultado da análise mostrou que a CNN obteve um desempenho melhor para detectar fraude do que outros modelos de Aprendizado de Máquina como Redes Neurais Artificiais, *Support Vector Regression* e *Random Forrest*.

Seguindo a mesma linha de estudo de Fu et al. (2016), Raza e Qayyum (2019) salientam que identificar transações fraudulentas em tempo real é um dos desafios

da atualidade, tendo em vista a quantidade crescente de *"cybercrime"* com particular ênfase em transações de cartão de crédito. No contexto de fraudes de cartão de crédito, a Detecção de Anomalia é um importante processo que auxilia na descoberta de padrões anômalos dentro de um *dataset* e como dados fraudulentos tem comportamento anômalo, o uso dessa abordagem nessa área é importante para detecção destes dados. Raza e Qayyum (2019) propõem a aplicação de *Variational Auto-Encoder* (VAE) para detecção de fraudes em cartão de crédito. Para efeito de comparação de resultados, utilizou os Modelos de Árvore de Decisão, Máquina de Suporte Vetorial e Classificadores *Ensemble*. Segundo as métricas de *recall* e precisão utilizas para avaliar os resultados, os modelos que apresentaram os melhores desempenhos foi Classificadores *Ensemble* (Adaboost) para precisão e, para *recall*, o melhor resultado foi do modelo VAE.

Dentro dos métodos de Aprendizado de Máquina, existem modelos de Aprendizagem Profunda, mais comumente chamados na literatura de *Deep Learning*. Os modelos de *Deep Learing* utilizam várias camadas de processamento para fazer *Feature Extration*[3] e *Feature Transformation*[4]. Estes algoritmos também são aplicados para dados de fraude financeira e possuem um bom desempenho segundo o estudo de Roy et al. (2018) que aplica 4 modelos de classificação de *Deep Learning* (Redes Neurais, Redes Neurais Recorrentes, *Long Short-term Memory* - LSTM e *Gated Recurrent Units* - GRU) em dados extremamente desbalanceados de transações de cartão de crédito, em que 0,14% dos dados era considerados como dados fraudulentos e 99,86% eram considerados normais. O estudo mostrou que LSTM e GRU obtiveram resultados superiores aos outros métodos de Rede Neural. Contudo, os autores sugeriam que os resultados das Redes Neurais poderiam ser aprimorados aumentando as camadas e neurônios da rede, entretanto ainda conforme os autores mesmo que o modelo gere melhores resultados há um custo computacional alto.

Saindo um pouco do escopo de fraude de cartão de crédito, Kirkos, Spathis e Manolopoulos (2007) realizou um profundo estudo para auxiliar o trabalho de auditoria em demonstrações financeiras fraudulentas na Grécia utilizando estatística e Inteligência Artificial. O estudo utiliza *Decision Trees*, Redes Neurais e *Bayesian Belief Networks* em uma base com dados de índices financeiros e demonstrativos financeiros de 76 empresas gregas. Os resultados são benéficos principalmente para os trabalhos de autoria

[3] *Feature Extration* é uma forma de redução de dimensionalidade que mapeia os recursos mais úteis dentro de todo o conteúdo de informações.

[4] *Feature Transformation* é o processo de transformação dos recursos de uma representação para outra.

e contabilidade na detecção de fraudes em demonstrativos financeiros, mas também traz auxílio para autoridades tributárias, governamentais e analistas econômicos. O modelo que obteve melhor resultado foi *Bayesian Belief Networks* seguido por *Decision Trees*.O estudo não focou com afinco na questão do desbalanceamento fazendo apenas alguns tratamentos de seleção de variáveis para reduzir a dimensionalidade.

Outra aplicação possível que tem muito potencial de estudos na área, é utilizar métodos de Aprendizado de Máquina para Detecção de Anomalia em *bitcoins*. O estudo Pham e Lee (2016) procurou detectar as transações ou usuários suspeitos da rede de Bitcoins. Bitcoins são um tipo diferenciado de transações financeiras utilizando "moedas virtuais" e tecnologia *peer-to-peer* (Bitcoin, 2019). Pham e Lee (2016) focaram em detectar anomalias em dois grafos gerados pela redes de transação de *Bitcoins* utilizando Métodos de Clusterização *K-means*, Método baseado em Distância Mahalanobis e Máquina de Suporte Vetorial Não-supervisionado. Um dos grafos tem as transações como nós e no outro grafo, os usurários como nós. O melhor resultado se deu no SVM Não supervisionado.

Fraudes bancárias também são alvo de aplicações de métodos de Aprendizado de Máquina. Abdelhamid, Khaoula e Atika (2014) estabelecem um análise do problema de fraude bancária utilizando variações de modelos de SVM. As fraudes bancárias podem ser externas, por instituições financeiras ou pessoas, e internas, por funcionários dos próprios bancos, e três são os focos do estudo citado cima: fraude de cartão de crédito, lavagem de dinheiro e fraude de hipoteca. O método aplicado é um modelo de SVM híbrido de aprendizado supervisionado e não supervisionado. A abordagem supervisionada tem como objetivo separar as transações fraudulentas das que não são e a abordagem não supervisionada tem como objetivo detectar as transações fraudulentas, os resultados do método apresentaram uma ligeira melhora em comparação com os métodos utilizados na literatura.

Como forma de viabilizar e facilitar as análises de Eventos Raros, King e Zeng (2001) propõem uma série de métodos como forma de melhorar as performances de modelos nos bancos de dados que possuem características de eventos raros binários, estes que ocorrem com uma frequência muito baixa, que é o caso de dados de fraude. A ideia aplicada no estudo de King e Zeng (2001) é importante para buscar um conjunto de métricas que auxiliem os modelos aplicados em dados de fraude e também em dados que possuem comportamento de evento raro, já que eventos deste tipo ocorrem em várias áreas do conhecimento, como na ciência política, estudos de

ocorrência de guerra, vetos presidenciais ou golpes, na medicina, com infecções de doenças incomuns e entre outros exemplos.

Mensah et al. (2019) apresenta em seu estudo a aplicação de três algoritmos para detectar operações fraudulentas em agências de viagem online. As três técnicas de Detecção de Anomalia Utilizadas foram *One-Class SVM*, *K-means clustering* e *Isolation Forrest*, como os dados não são rotulados, os autores assumiram que seria melhor utilizar métodos não-supervisionados.

Os estudos citados mostram que existe uma tendência de uso dos modelos de Aprendizado de Máquina para detectar ou prever fraudes (RAZA; QAYYUM, 2019). Contudo, não há muitas aplicações de AM para previsão de Fraudes em Fundos de Pensão, principalmente no caso brasileiro. Nas próximas seções, serão apresentados alguns exemplos de Modelos Supervisionados e Não-Supervisionados de AM utilizados para Eventos Raros.

2.2 Modelos de Aprendizado de Máquina para Eventos Raros

Modelos de Aprendizado de Máquina (AM) têm sido amplamente utilizados para detecção de fraudes e em eventos raros em geral (RAZA; QAYYUM, 2019; KAISER et al., 2017). Os estudos na área de AM em Eventos Raros são importantes pelos seguintes motivos: primeiro porque a maioria dos modelos e algoritmos supõe que as classes de dados já são balanceadas e distribuídas uniformemente, enquanto que os dados reais não se comportam dessa maneira. Logo para que os modelos realmente consigam interpretar a realidade, é preciso lidar com problemas que possuem a característica de eventos raros. Além disso, as aplicações de dados de eventos raros na vida real são muito importantes de serem estudados porque as previsões desse tipo de eventos são primordiais para evitar consequências ruins em diversas áreas como financeira (fraude ou falência) e até climática (tempestades, furações, etc.) (MAALOUF; SIDDIQI, 2014). Esta Seção expõe alguns métodos de AM utilizados para detectar ou prever eventos raros em finanças e também em outras áreas.

Para modelos de classificação binária, são considerados Eventos Raros, dados em que as variáveis dependentes possuem uma quantidade muito baixa do evento focal (rotulado usualmente com o valor numérico um), que representa a ocorrência do evento, em comparação à ausência do evento (rotulado numericamente com o valor zero) (JANJUA et al., 2019; KING; ZENG, 2001; MAALOUF; TRAFALIS, 2011). Em dados reais, é

muito comum encontrar este desequilíbrio das classes de ocorrência em campos como medicina (pacientes saudáveis e não saudáveis), segurança em redes de computadores (atividade não maliciosa ou maliciosa), visão computacional (objetivo da visão ou pano de fundo) e fraude (atividade comum ou fraudulenta) (KRAWCZYK, 2016). Neste cenário ocorre sempre uma classe majoritária e uma classe minoritária e existem técnicas na tentativa de equilibrar o balanceamento dessas bases de dados. Segundo Krawczyk (2016), classificação em dados desbalanceados para multi-classe não é tão bem desenvolvida como para classificação binária, principalmente pela dificuldade de definição de relação entre as classes, isto é, uma classe pode ser majoritária em relação a outra entretanto minoritária em relação a outra.

Outra abordagem possível em estudos de Eventos Raros ocorre em modelos do tipo Regressão, isto é, fazendo detecção ou previsão de dados contínuos. Assim como para classificação binária, para modelos de regressão também há muitas aplicações em dados reais, como exemplo é possível citar gerenciamento de crise, estimação de perdas em finanças e previsão de temperatura em meteorologia (KRAWCZYK, 2016; TORGO et al., 2015). No caso destas análises, o objetivo é prever prováveis valores extremos nos conjuntos de dados. Estes valores são chamados de *outliers* e são estudados pela vertente da Estatística que foca em análises de valores anormalmente altos ou baixos, denominada Teoria dos Valores Extremos (TORGO et al., 2015; HAIXIANG et al., 2017).

No presente estudo, o foco está em Modelos de Classificação Supervisionados e em Técnicas de Balanceamento para prever eventos raros que, neste caso, são as Fraudes em Fundos de Pensão. Em várias áreas do conhecimento é possível encontrar exemplos de eventos raros e extremos, como em ciência política, nos estudos de ocorrência de guerra, vetos presidenciais ou golpes, ou em medicina, com infecções de doenças incomuns (KING; ZENG, 2001; MAALOUF; TRAFALIS, 2011). A primeira dificuldade apontada por King e Zeng (2001) é a aplicação de modelo de Regressão Logística em dados de eventos raros, porque tende a trazer resultados enviesados para as amostras pequenas que representam esses eventos. Por mais que o método seja quase universalmente utilizado por estudiosos da área, sem as adaptações corretas, pode trazer resultados que não condizem com a realidade, demonstrando um desempenho sub-ótimo para prever eventos raros.

A segunda dificuldade apontada remete à coleta de dados em eventos raros, principalmente aos critérios usados para adicionar conjuntos de dados com poucas variá-

veis explicativas, que podem agregar mais ao modelo. Umas das principais aplicações do estudo de eventos extremos são em pesquisa de Eventos Naturais, como precipitações, inundações, correntes oceânicas, poluição atmosférica, principalmente para aplicações em modelos de regressão, entretanto com o avanço dos estudos novas aplicações foram sendo acrescentadas em modelos de classificação nas áreas de engenharia, finanças e atuária (MENDES, 2004).

Na Literatura de Aprendizado de Máquina é evidenciado alguns problemas associados a detecção ou previsão de eventos raros. Segundo Weiss (2004), as dificuldades mais comuns na aplicação deste tipo de modelos são:

1. Uso de Métricas de Avaliação Inapropriadas: As métricas de avaliação são utilizadas para verificar o desempenho dos algoritmos. Métricas utilizadas pata modelos gerais, a mais comum delas sendo a acurácia, não são tão boas para avaliar modelos que procuram detectar eventos raros porque dão viés para as classes majoritárias dos dados.

2. Falta de Dados (Raridade Absoluta ou Raridade Relativa): Uma dificuldade fundamental para lidar com eventos raros é falta de dados, tanto no sentido absoluto, quando no número de casos relacionados a uma determinada classe é muito menor que as outras classes dos dados, como no sentido relativo, quando os dados são difíceis de detectar pelos próprios modelos.

3. Fragmentação de Dados: a questão da fragmentação dos dados está relacionada com os algoritmos que possuem abordagem de dividir e conquistar, como *Decision Trees* por exemplo. A ideia principal desta abordagem é particionar os dados em pedaços menores, a dificuldade envolvida nestas abordagens é que os possíveis padrões encontrados nas partições são encontrados somente nessas, de forma individual, o que prejudica a identificação de padrões de eventos raros.

4. Viés Indutivo Inapropriado: neste caso, a dificuldade se encontra nos modelos de generalização que quando buscam generalidade máxima, desconsideram os eventos raros do conjunto de dados por considerar os conjuntos majoritários dos dados.

5. Ruído: ruídos são dados que possuem comportamentos inconsistentes com o conjunto de dados na qual pertencem, por ter esse tipo de comportamento po-

dem ser confundidos com eventos raros que são o foco da detecção, trazendo dificuldade para os modelos.

A seguir, será tratado acerca de modelos de Aprendizado de Máquina que são utilizados para detectar ou prever eventos raros com foco principal em finanças e administração.

2.2.1 Modelos Supervisionados para Eventos Raros

2.2.1.1 Regressão Logística

A Regressão Logística (RL) é um dos classificadores de aprendizado supervisionado mais clássicos da literatura da área, sendo usado em diversas segmentos como mineração de dados, medicina e economia (ZHANG et al., 2019; MAALOUF; HOMOUZ; TRAFALIS, 2018; KING; ZENG, 2001). O resultado do modelo de regressão logística são as previsões de probabilidades das classes dos dados, o limite padrão para este tipo de modelos geralmente é $0,5$. Os estudos de Zhang et al. (2019) tentaram ajustar este valor padrão para tentar se adequar ao um conjunto de dados desbalanceados, entretanto ao ajustar o limite padrão, a acurácia do modelo foi afetada.

King e Zeng (2001) propõe ajustes que julgam necessários na aplicação de Regressão Logística em eventos raros. Um dos problemas principais elencados pelos autores nos Modelos de Regressão Logística está na estimação das probabilidades dos eventos, porque em caso de amostras finitas de eventos raros estas probabilidades se tornam sub-ótimas trazendo resultados enviesados. A sugestão de King e Zeng (2001) para solucionar este problema é no uso de correções nos parâmetros da Regressão, com a utilização de pesos para corrigir o viés, e consequentemente, o ajuste no cálculos das probabilidades. O modelo proposto por King e Zeng (2001) não foi aplicado na área de Finanças, os autores recomendaram o uso do modelo no Estudo de Conflitos Internacionais.

Outra abordagem que pode contribuir para aumentar os níveis de acurácia dos modelos de Regressão Logística é o método SMOTE, que é um método *over-sampling* (ou “super-amostragem”) que adiciona dados sintéticos com base nos dados originais para “balancear” o conjunto de dados (RAHIM et al., 2019; KÖKNAR-TEZEL; LATECKI, 2009). Em Regressão Logística, o algoritmo SMOTE foi aplicado em um estudo feito na tentativa de prever falência de médias e pequenas empresas e apresentou uma melhora

significativa na acurácia em comparação a um modelo de RL sem a técnica de amostragem (RAHIM et al., 2019). Outras técnicas de amostragem que também podem ser usadas para superar os problemas causados por dados desbalanceados são: BLSMOTE, MWMOTE e KMSMOTE.

O *Borderline-SMOTE* (BLSMOTE) é um método de *oversampling* que faz "super-amostragem" apenas dos chamados exemplos de fronteiras das classes minoritárias. Exemplos de fronteiras são os dados de minorias limítrofes, que segundo Han, Wang e Mao (2005) são os valores que influenciam mais nos erros de classificação. A grande diferença de BLSMOTE para o SMOTE é o uso dos exemplos de fronteira para a super-amostragem, no modelo tradicional SMOTE apenas amplia os dados nas classes minoritárias (SARKAR et al., 2020).

A ideia principal do *Majority Weighted Minority Oversampling Technique* (MWMOTE) é criar amostras das classes minoritárias e atribuir pesos de acordo com a distância euclidiana a partir das amostras das classes majoritárias mais próximas (SARKAR et al., 2020). O método realiza os seguintes passos: seleção de amostras da classe minoritária, atribuição de pesos às amostras classificadas como mais importantes e agrupamento para gerar a composição sintética das amostras da classe. O objetivo principal do MWMOTE é melhorar a maneira em que se seleciona as sub-amostras e na geração das amostras sintéticas (BARUA et al., 2012).

O *K-means* SMOTE (KMSMOTE) é uma união dos algoritmos *K-means clustering* com a técnica SMOTE. O *K-means clustering* permite que método identifique em quais grupos dos dados de entrada será mais eficaz para a geração dos dados artificiais. O objetivo não é apenas melhorar o desbalanceamento das bases com a geração dos dados das classes minoritárias, mas também melhorar os dados dentro da classe, evitando geração de ruídos. A diminuição da geração de ruídos se dá por meio do uso da super-amostragem do SMOTE somente nos agrupamentos relevantes para geração dos dados, isto é, as classes minoritárias. A escolha dos agrupamentos que serão usados o SMOTE é baseado na proporção de classes minoritárias e majoritárias de cada um dos grupos (DOUZAS; BACAO; LAST, 2018)

2.2.1.2 Árvore de Decisão

Modelos de Árvore de Decisão (*Decision Trees*) utilizam estrutura de árvore para separar os dados em subgrupos. A estrutura cria nós de forma binária ou múltipla, a

partir de um nós raiz, que separa os dados de acordo com os grupos (KIRKOS; SPATHIS; MANOLOPOULOS, 2007). O processo recursivo da árvore só termina quando todos os dados de um nós pertencem a uma única classe ou quando não há mais recursos para expandir a árvore e, geralmente, os critérios de parada são definidos para impedir *overfitting* (LI, 2007). Em eventos raros, uma maneira de melhorar o desempenho e a acurácia desses modelos dada as dificuldades já citadas é trabalhar na estrutura da árvore ou utilizar Classificadores *Ensemble* (GALAR et al., 2011). Além disso, o estudo de Lee (2019) propôs uma maneira de melhorar a acurácia de Modelos de Árvore de Decisão em dados desbalanceados, usando a própria acurácia para escolha do melhor atributo de divisão da árvore. Comparando o modelo propostos com Árvore de Decisão pura e também com Árvore de Decisão com aprendizado sensível a custo, o modelo proposto obteve melhores resultados.

2.2.1.3 *Random Forest*

Random Forest é um método de aprendizado para classificação ou regressão que se baseia em uma coleção de árvores de decisão $\{h(\mathbf{x}, \Theta_k), k = 1, ...\}$ em que os $\{\Theta_k)\}$ é o vetor (ou matrix) de parâmetros desconhecidos e $\mathbf{x}$ são vetores de entrada, aleatórios independente e identicamente distribuídos, e cada árvore gera um voto unitário para a classe mais popular dos dados de entrada x. O modelo é formado por árvores de decisão que crescem a partir dos dados de entrada ou de combinações de entradas, estes devem ser selecionados aleatoriamente com uso de *bagging*, que consiste em uma metodologia de amostragem aleatória para substituir as combinações de entrada dos dados de treinamento, ou somente com uso de *bootstrap*.

Há duas razões principais para o uso de *bagging* em *Random Forest*, em primeiro lugar o método aumenta a acurácia dos resultados e, além disso, fornece estimativas dos erros de generalização de forma contínua das árvores de decisão (BREIMAN, 2001).

Para dados desbalanceados, um estudo feito por Muchlinski et al. (2016) comparando a performance de *Random Forest* e Regressão Logística para prever eventos de guerra civil, um tema de suma importância para estudos preditivos em Ciência Política. Neste estudo, *Random Forest* teve um poder preditivo superior que as variações de Regressão Logística e apresentou uma maior flexibilidade para previsões mais precisas deste tipo de evento (MUCHLINSKI et al., 2016).

O modelo original de *Random Forest* proposto por Breiman (2001) funciona bem

em várias aplicações para finanças(KRAUSS; DO; HUCK, 2017; KUMAR; THENMOZHI, 2014). Entretanto para dados desbalanceados, Chen et al. (2004) propôs duas adaptações ao modelo: *Balanced Ramdon Forest* (*Random Forest* Balanceado - BRF) e *Weighted Random Forest* (*Random Forest Ponderado)*. O primeiro modelo estratifica a amostra para que os dados tenham categorias mais "balanceadas" , aumentando a ocorrência das categorias minoritárias. O segundo modelo confere pesos maiores de erro para as categorias mais raras (CHEN et al., 2004; EFFENDY; BAIZAL et al., 2014).

2.2.1.4 Classificadores *Ensemble*.

O objetivo principal dos Métodos *Ensemble* é aprimorar o desempenho de modelos de classificação, combinando vários classificadores para obter um modelo que supere os demais. A motivação de combinar classificadores é melhorar a capacidade de generalização e tentar diminuir os erros gerados nos modelos, por isso são muito utilizados em bases de dados desbalanceados para apurar modelos de classificação (GALAR et al., 2011). Dois exemplos de Métodos serão expostos a seguir: *Bagging* e *Boosting*. *Bagging* (*Bootstrap Aggregating*) gera amostras dos dados originais por Amostragem *Bootstrap*. Os conjuntos de dados gerados são treinados por modelos de classificação e os resultados dos modelos são dados por votos das classes majoritárias, ou seja, o resultado se dá pelo conjunto das análises (BREIMAN, 1996; GALAR et al., 2011). No *Boosting*, semelhantemente ao *Bagging*, também utiliza diferentes conjuntos de dados de treinamento, entretanto o algoritmo recalcula os resultados baseados nos acertos dos votos anteriores, ou seja, os pesos são remodelados em cada conjunto de treinamento e esse procedimento garante um aumento de desempenho dos classificadores, o algoritmo mais famoso desta classe é o *AdaBoost* (GALAR et al., 2011).

2.2.1.5 *Gradient Boosting*

Gradient Boosting são um grupo de modelos que tem sido muito utilizado em técnicas de Aprendizado de Máquina principalmente pelo surgimento de novos estudos na área de tecnologia e *big data* (KE et al., 2017). A lógica dos modelos *Gradient Boosting* é a combinação de vários modelos "fracos". O processo de aprendizado vai se ajustando a medida que os modelos são adicionados e o objetivo é produzir uma estimativa mais precisa das variável resposta. Os modelos são construídos em etapas e em cada etapa, os modelos são correlacionados com um gradiente negativo da função de perda. A função de perda é uma medida de ajuste dos coeficientes do mo-

delo e é utilizada para que nas etapa do *Gradient Boosting* os erros sejam minimizados (FRIEDMAN, 2001; NATEKIN; KNOLL, 2013)

Algoritmos de *Gradient Boosting* utilizam da lógica de Aprendizado por Árvore de Decisão para modelagem preditiva e estes modelos precisam estudar todas as instâncias das *features* para estimar qual seria o ganho de informação em cada nó de decisão. Isso mostra que a complexidade computacional destes modelos depende da quantidade de dados e dos recursos das bases de dados, logo os modelos com implementações muito demoradas devido ao uso de bases com muitos dados e com um alto uso de memória nos sistemas computacionais. Dois exemplos deste tipo de algoritmo é *Extreme Gradient Boosting*, ou XGBoost e *Light Gradient Boosting Model*, ou LightGBM.

Da mesma forma que o LightGBM, O XGBoost ou *Extreme Gradient Boosting*, também é um algoritmo de Aprendizado de Máquina baseado em árvore de decisão que se baseia em *Gradient Boosting Framework*. O Modelo pode ser usado para classificação ou regressão. o algoritmo otimiza certos parâmetros, como função de perda, do aprendizado em Árvore, fazendo com o XGBoost seja mais rápido e tenha mais acurácia que outros modelos. De acordo com Chen e Guestrin (2016), o principal diferencial do XGboost é a possibilidade do algoritmo de Aprendizado de Máquina ser escalável a vários cenários, isso se dá porque o modelo consegue lidar dos dados esparsos e a um procedimento de esboço de quartil ponderado que permite a manipulação de pesos de instância no aprendizado aproximado das árvores. Além disso, a computação paralela e distribuída auxilia na rapidez e na exploração de várias modelagens para os dados.

O algoritmo tem sido muito utilizado na plataforma *Kaggle* de competições em Ciência de Dados e tem obtido êxito na maioria das vezes que é utilizado.Em 2015, de 29 soluções vencedoras em competições do *Kaggle*, 17 foram utilizando XGBoost. No estudo de fraudes, o modelo foi utilizado no estudo de Albiero et al. (2019) para prever irregularidades em consumo de energia elétrica. O estudo mostrou que o modelo de XGBoost apresentou melhores resultados que Regressão Logística para as métricas de *F1-Score*, *Precision* e *Recall*.

Outro estudo que utilizou XGBoost para prever fraudes foi feito por Zhang et al. (2020) que tem por objetivo detectar fraudes em transações de crédito de clientes utilizando uma base de dados pública da UCI com informações de transações de crédito feitos na Alemanha. Além de XGBoost, foram usados outros métodos de Aprendizado

de Máquina como *benchmark* e são Regressão Logística, *Random Forest* e *Support Vector Machine*. Entre estes modelos, o XGBoost obteve melhor desempenho pelas métricas de Curva AUC e Acurácia.

O LightGBM surgiu com a proposta de solucionar os problemas de com tempo de processamento e com o uso da memória. O Modelo propõe o uso de duas técnicas para aprimorar os algoritmos de *Gradient Boosting*: Amostragem unilateral baseado em Gradiente (*Gradient-based One-Side Sampling* - GOSS) e Pacote de Recursos Exclusivos (*Exclusive Feature Bundling* - EFB). Algoritmos de *Gradient Boosting* que utilizam GOSS e EFB são denominados de *Light Gradient Boosting Models*, ou LightGBM. Este modelo se tornou muito popular em estudos e em competições na plataforma de competições de Ciência de Dados *Kaggle*, principalmente pela velocidade de processamento, menor uso de memória e melhor acurácia das previsões (KE et al., 2017).

Dentro do contexto de Fraudes, a aplicação de LightGBM apresenta bons resultados, em comparação outros modelos também utilizados para precisão (TAHA; MALEBARY, 2020; HU; CHEN; ZHANG, 2019). O modelo é bastante aplicado para Fraude em Cartão de Crédito. No estudo de Taha e Malebary (2020), é proposto um modelo de LightGBM Otimizado para prever Fraude em Cartão de Crédito. A abordagem é feita com um algoritmo de otimização bayesiana nos hiperparâmetros do modelo para melhorar a performance. São utilizadas duas bases de dados públicas e e foram utilizadas 5 métricas de avaliação: Acurácia, *Recall*, AUC, Precisão e *F1-Score*. O LightGBM otimizado apresentou melhores números para Acurácia, AUC, Precisão e *F1 - Score* em comparação a outros modelos de Aprendizado de Máquina, que são *Random Forest*, Regressão Logística, Máquina de Suporte Vetorial Radial, Máquina de Suporte Vetorial Linear, *K-Nearest Neighbors*, Árvore de Decisão e *Naive Bayes*.

2.2.1.6 Máquina de Suporte Vetorial

Máquina de Suporte Vetorial (*Support Vector Machine* - SVM) são algoritmos de Aprendizado de Máquina que buscam generalizar uma função utilizando as chamadas função *kernel*, ou seja, o SVM procura aprender como os modelos ocorrem e generalizar os resultados de dentro da amostra de treinamento para todo o conjunto de dados. Em suma, os SVMs são classificadores lineares que fazem mapeamento não linear do espaço de entrada em um espaço de alta dimensão (BHATTACHARYYA et al., 2011). Trabalhar com espaço de alta dimensão é uma vantagem para o modelo por conseguir flexibilizar a aplicação para vários tipos de cenários de dados. O SVM

mais simples de ser aplicado é o que busca obter a melhor margem para separação de dois grupos de dados, de modo a buscar a máxima separação entre duas classes distintas, no caso de dados binários (RAZA; QAYYUM, 2019). O primeiro modelo de SVM foi construído por Cortes e Vapnik (1995) na tentativa de solucionar de maneira otimizada modelos de classificação e regressão. Modelos de SVM têm propriedades que trabalham bem dados lineares e não lineares, fazendo com que suas aplicações sejam bem versáteis em diversas áreas, principalmente em estudos de classificação binária, como é o caso de precisão de fraude (BHATTACHARYYA et al., 2011).

Em eventos raros, SVM tem melhor desempenho quando utilizado em junção de outros algoritmos para melhorar acurácia como é o exemplo do estudo de Chi e Ersoy (2002) que aprimorou um modelo de Máquina de Suporte Vetorial juntando com um modelo de Árvore de Decisão chamado LSVM-DT, o modelo utiliza um SVM linear em cada nó da Árvore de Decisão. A união dos dois modelos possibilita uma estrutura que melhora a performance de ambos os modelos e permitem a detecção de eventos raros. Outras duas abordagens que podem auxiliar a detecção de eventos raros são as amostragens aleatórias e o Método SMOTE (Synthetic Minority Over-sampling Technique) que utilizam métodos de amostragem para remoção de dados das classes majoritárias com objetivo de melhorar a acurácia de previsão das classes minoritárias (eventos raros). A amostragem aleatória replica os dados raros do próprio conjunto de dados de maneira aleatória, a técnica pode até ser efetiva, porém há o perigo de *overfitting*[5]. O Método SMOTE gera novos dados simulados das classes minoritárias e, por isso, demonstra uma melhora na precisão geral e no aprendizado dos modelos. (KÖKNAR-TEZEL; LATECKI, 2009). Utilizando essa mesma lógica de adicionar novos dados ao conjunto total, o estudo de Köknar-Tezel e Latecki (2009) propôs a adição de *ghost points* no conjunto de dados de distâncias espaciais para melhorar a classificação de um modelo de Máquina de Suporte Vetorial e obteve uma acurácia significativamente maior que nos modelos de adição de novos dados.

2.2.1.7 Redes Neurais

Redes Neurais Artificiais (*Artificial Neural Networks - ANN*), ou apenas Redes Neurais, são modelos estatísticos não-paramétricos composto por várias unidades de processamento, também chamados de neurônios, que armazenam as informações como forma de conhecimento para ser utilizado pelo modelo posteriormente com a

[5] *Overfitting* ocorre quando um modelo estatístico ajusta muito aos dados de treinamento e não é eficiente para fazer novas previsões, ou seja, não generaliza bem para novos dados

devida ponderação necessária (WEST; BHATTACHARYA, 2016; HAYKIN, 1994). A complexidade computacional de uma Rede Neural depende muito de sua estrutura e de sua capacidade de generalizar (HAYKIN, 1994). Em eventos raros, o modelo é muito utilizado para detecção de ruídos de sons, detecção de ruídos em imagem e fraude em transações de cartão de crédito (CAKIR; VIRTANEN, 2019; WANG; KAO; WANG, 2018; MAES et al., 2002). No caso de fraudes, as redes neurais puras tem um desempeno ruim e, em estudos mais recentes têm sido mostrado que as Redes Neurais Convolucionais (Convolutional Neural Network - CNN) possuem um desempenho maior na detecção destes eventos (ZHANG et al., 2018; FU et al., 2016). CNN são modelos que são bons para treinamento de conjuntos com muitos dados e possuem mecanismos para evitar *overfitting* (FU et al., 2016).

Raj, Magg e Wermter (2016) apresenta soluções para a dificuldade de aplicação dos modelos de Rede Neural em dados desbalanceados. Os estudo propõe diferentes métodos para criar uma abordagem de pesos para separabilidade das classes utilizando Redes Neurais Convolucionais sensíveis a Custos. Segundo Raj, Magg e Wermter (2016), a principal dificuldade de classificação dos modelos em dados desbalanceados está na separabilidade das classes. Utilizando a técnica de Erro Global Médio de Separação (*Global Mean Error Separation* - GMSE), desenvolvido por Castro e Braga (2009), para realizar a aprendizagem sensível a custo com a otimização de um parâmetro para penalizar o erro de classificação das classes majoritárias. Com os ajustes realizados, a CNN sensível a custos obteve melhores resultados que a CNN sem a adaptação. A métrica de avaliação utilizada para verificar o desempenho do modelos foi *G-Mean Score*.

Dentro desta subseção foram listadas algumas aplicações de Redes Neurais dentro da área de Finanças, de Eventos Raros em Finanças e Dados desbalanceados. Entretanto ainda há uma vasta literatura e aplicações de Redes Neurais em Eventos Raros em outros campos de estudo.

3 METODOLOGIA

Nas seções a seguir, é descrita a análise qualitativa inicial realizada para verificar quais são as variáveis importantes para previsão de indício de fraude, os dados utilizados no estudo, a seleção de variáveis realizada, os métodos utilizados na seleção bem como as métricas que foram consideradas nas avaliação de desempenho dos modelos escolhidos aplicados diretamente na Base de Dados da Superintendência Nacional de Previdência Complementar (PREVIC).

O modelo de Regressão Logística foi considerado o modelo base de comparação (Modelo *Baseline*), *Random Forest*, Máquina de Suporte Vetorial, *LightGBM* e Redes Neurais foram comparados segundo o modelo *Baseline*. Dentro de cada modelo, o desempenho dos métodos foram comparados para se verificar qual o melhor modo de prever o Indício de Fraudes em Fundo de Pensão.

Para diferentes tipos de desbalanceamento de uma base de dados, a previsão pode se comportar de diferentes formas, devido à falta de representação dos dados nas classes minoritárias. Isso significa que para bases que possuem um balanceamento de 50% entre as classes, os modelos de previsão podem apresentar resultados melhores em comparação à bases de dados que, por exemplo, possuem um desbalanceamento de 5% de uma classe para 95% de outra classe pelos motivos de: uso de métricas inadequadas, falta de dados nas classes minoritárias, fragmentação de dados, viés indutivo inapropriado além de se considerar as classes minoritárias como ruídos nos dados. Os classificadores geralmente enviesam a previsão para as classes com maior parte dos dados e apresentam taxas de classificação mais baixas para as classes minoritárias. Para verificar o comportamento das métricas para bases desbalanceadas, foi realizado um estudo com simulação de dados para analisar como as métricas de desempenho dos modelos se comportam segundo diferentes proporções de desbalanceamento e tamanhos de amostra.

3.1 Análise Qualitativa

Uma das etapas realizadas na Metodologia deste estudo foi Análise Qualitativa utilizando a Experiência dos Auditores da PREVIC. Foram realizadas 4 entrevistas nos dias 07, 14, 21 e 29 do mês de Julho de 2020 com 3 Servidores que fazem parte do Grupo de Trabalho do Estudo de Fraudes que possuem Cargo de Auditoria para estabelecer um padrão de funcionamento dos casos de fraude e verificar quais variáveis seriam mais relevantes para utilizar no trabalho.

Com as entrevistas, percebeu-se que o objetivo da PREVIC é verificar o indício da fraude para indicar que seja feita uma investigação mais profunda por meio do Ministério Público e Polícia Federal. Isso acontece porque a empresa não possui poder de punição, de acordo com a portaria que criou o órgão. Os resultados dos Modelos apresentados neste estudo indicam uma direção nas investigações da PREVIC com a detecção de possíveis Entidades com suspeita de Fraudes. A análise realizada pelos Auditores é puramente manual e com grande quantidade de estudo humano no processo, observando os padrões Financeiros do Fundo, isto é, os padrões da carteira de ativos financeiros, observando o comportamento dos Dirigentes dos Fundos e observando as características principais dos Fundos de Pensão.

3.2 Base de Dados

Nessa Seção será abordada a operacionalização dos dados que formaram a base de dados do modelo. Os dados são formados por três dimensões: Dimensão do Fundo de Pensão, que apresenta características dos fundos; Dimensão do Dirigente do Fundo, com características dos tomadores de decisão dos Fundos; e Dimensão de Investimentos, com informações acerca dos ativos que fazem parte da carteira do fundo. Além disso, as informações se encontram entre os anos de 2014 a 2019 e distribuídas mensalmente, onde em cada linha há informações com respeito ao fundo no mês.

Na Dimensão de Fundo de Pensão ou Entidade, as variáveis são: Idade do Fundo em meses no tempo e Patrocinador Predominante. A Idade do Fundo em meses é calculada pela diferença entre a Data de Fundação e o período analisado, isso significa que em cada linha há a quantidade de meses de existência do Fundo referente ao mês analisado na linha.

O Patrocinador Predominante se refere ao tipo de Patrocinador do Fundo. Patrocinador é empresa, grupo de empresas, União, Estados, o Distrito Federal, os Municípios ou outras entidades públicas que instituem os Fundos de caráter previdenciário para os seus colaboradores [1]. Esta variável por três colunas que se comportam como *dummy* [2] e recebem o valor "1" quando o Fundo tem como predominante um destes tipos de patrocinador: Instituidor, Privado, Público Federal, Pública Estadual e Pública Municipal. Os dados das Entidades foram fornecidos pela PREVIC. A Média de Idade dos Fundos é de 292 meses (24 anos) e metade (56%) deles são de Instituição Privada.

Na Dimensão dos Dirigentes, as variáveis de características dos Diretores dos Fundos são filiação partidária. De acordo com as Análise Qualitativa nas entrevistas realizadas com os servidores da PREVIC, o antecedente político do Dirigente é relevante para a análise, devido ao histórico de influência política em fraudes já ocorridas.

As informações acerca das filiações do partidos políticos foram retiradas da Base de Filiação Partidária do Tribunal Superior Eleitoral (TSE). A base foi cruzada com a Base de Dirigentes da PREVIC para obter a informação de quais dirigentes eram afiliados a alguns destes Partidos Políticos: Partido dos Trabalhadores (PT), Avante (AVANTE), Cidadania (CIDADANIA), Movimento Democrático Brasileiro (MDB), Partido Comunista Brasileiro (PCB), Partido Comunista do Brasil, Partido da Mobilização Nacional (PMN), Partido da Mulher Brasileira (PMB), Partido da Social Democracia Brasileira (PSDB) , Partido Democrático Trabalhista (PDT), Partido Liberal (PL), Partido Novo (NOVO), Partido Renovador Trabalhista Brasileiro (PRTB), Partido Republicano da Ordem Social (PROS), Partido Social Cristão (PSC), Partido Social Democrático (PSD), Partido Social Liberal (PSL), Partido Socialismo e Liberdade (PSOL), Partido Socialista dos Trabalhadores Unificado (PSTU), Partido Socialista Brasileiro (PSB), Partido Trabalhista Brasileiro (PTB), Partido Trabalhista Cristão (PTC) e Partido Verde (PV).

Considerando o tempo de identificação das fraudes, que segundo os servidores é de 5 anos devido a prescrição do Auto de Infração e do período de tempo que um Dirigente permanece no processo decisório de uma Entidade, a quantidade de afiliados por Entidade foi defasada em 12, 24, 36, 48 e 60 meses para se obter o número de Dirigente Filiados ao partido com uma defasagem de 1 a 5 anos antes

[1] http://www.previc.gov.br/a-previdencia-complementar-fechada/patrocinador-participante-e-assistido

[2] Variável *Dummy* é uma variável binária que é utilizada para representar variáveis categóricas

do referido tempo e, assim, criar as variáveis de acordo com defasagem temporal necessária de análise, tendo em vista que 5 anos é o período de prescrição do Auto de Infração. Alguns partidos que apresentaram maior quantidade de dirigentes que outros partidos, com foco para o PT (10%), PSDB (7,5%) e MDB(6%), os outros partidos possuem menos de 1% de ocorrência de dirigente filiados.

As variáveis que definem a Dimensão de Investimentos são formadas pelos retornos gerados de cada ativo na carteira do Fundo para avaliar possíveis mudanças em Política de Investimento. Para cada ativo que pertencente à Carteira de Investimentos, foi calculada a porcentagem de cada ativo da carteira baseado no valor absoluto do ativo em comparação ao valor total.

Após isso, foram calculados os retornos mensais defasados em 12, 24, 36, 48 e 60 meses para se obter o retorno dos ativos com a de 1 a 5 anos antes do referido tempo. Isso significa que, por exemplo, para o ativo de um fundo em um específico mês, haverá variáveis dos retornos deste ativo dos últimos 5 anos.

O objetivo de usar os retornos dos últimos 5 anos é para observar o comportamento do ativo ao longo do período de prescrição de um Auto de Infração. Além das variáveis relacionadas aos ativos financeiros das carteiras, também foi calculada uma variável para descrever a diferença entre a Data de Vencimento e o período de referência.

A variável de período ponderado consiste em uma ponderação entre as diferenças mensais da Data de Vencimento para o período de referência ponderadas pela quantidade dos ativos dentro da carteira de investimento. Esta variável tem objetivo de mensurar Liquidez dos Ativos dentro da carteira. Dentro da Dimensão de Investimentos, encontram-se as variáveis de Ações, Depósito, Direito Creditório, Empréstimo, Financiamento Imobiliário, Imóvel, Cotas de Fundos, Valores a Pagar e Receber, Operação Compromissada, Títulos Públicos e Títulos Privados. Dentre estes, as variáveis que apresentam maior concentração são Ações, Depósitos e Cotas de Fundos.

Além das dimensões citadas, a última variável que compõe a base de dados é a *dummy* para a ocorrência ou não de Auto de Infração para o fundo o que gerou um total de 235 variáveis e 17329 linhas, sendo cada linha representando um fundo no tempo. Há 95 ocorrências de Auto de Infração para as Entidades presentes nos dados o que gera um desbalanceamento de 99.5% de não ocorrência da fraude para, 0.5% para a ocorrência de fraude. Como evidenciado pela baixa porcentagem de ocorrência de fraude, o estudo se caracteriza como evento raro. Com vistas a auxiliar na abordagem

de detecção deste evento, que tende a ser com análise manual feita pelos Auditores da PREVIC, e também a evitar prejuízos financeiros que podem ser causados, faz-se necessário a detecção do indício de Fraude em Fundos de Pensão.

3.3 Seleção de Variáveis

3.3.1 Regressão LASSO e Regressão Ridge

Tendo em vista a grande quantidade de variáveis presente na base de dados, foi-se necessário utilizar algumas técnicas de *Feature Selection* a fim de selecionar as variáveis que conseguissem explicar melhor as informações da base. Como primeira abordagem, foram retiradas da análise as variáveis que tinham uma variância menor que 0.00001 com isso a quantidade de variáveis de 235 para 87. A partir desta quantidade de variáveis encontrada após o filtro da variância, foi aplicada as técnicas Regressão LASSO (do inglês, *Least Absolute Shrinkage and Selection Operator*) e Regressão *Ridge*.

LASSO e Regressão Ridge são baseadas em Métodos de ”Encolhimento“ (*Shrinkage Methods* em inglês). Estes métodos minimizam a soma residual dos quadrados do modelos utilizando Mínimos Quadrados Ordinários para estimar os betas e também reduzem a complexidade do modelo, isso significa diminuir o número de preditores, com objetivo de reduzir a dimensão, selecionando um subconjunto relevante de dimensões. A princípio, a descrição do métodos parece bem semelhante a uma Regressão Linear comum, entretanto para Regressão LASSO a diferença principal é que o modelo adiciona uma penalidade para o coeficientes que são diferentes de zero, penalizando a soma dos valores absolutos chamado de penalização L1. Isto significa que, a LASSO transforma os coeficientes em componentes λ e truncando em zero e após isso, é subtraído a soma residual dos quadrados sujeito à soma do valor absoluto dos coeficientes. Na Equação a seguir, é possível matematicamente como funciona o modelo para estimação dos parâmetros:

$$l_{\lambda}^{L}(\beta) = \sum_{i=1}^{n} [y_i x_i \beta - log(1 + e^{x_i \beta})] - \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.1}$$

A regressão Ridge, também possui uma lógica de penalização para coeficientes. A diferença da Regressão Ridge está na utilização da penalização L2 que consiste no quadrado nos coeficientes ao invés do valor absoluto como é na Regressão LASSO. É possível observar como o modelo funciona para estimação dos parâmetros mate-

maticamente por meio da equação a seguir:

$$l_{\lambda}^{R}(\beta) = \sum_{i=1}^{n}[y_i x_i \beta - log(1 + e^{x_i\beta})] - \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.2}$$

Mesmo com o uso da Regressão LASSO e da Regressão Ridge, os resultados não se mostraram satisfatórios pois os betas dos coeficientes para ambas as regressões não foram representativos e por isso não foi possível estabelecer quais seriam as variáveis mais relevantes. Na Regressão LASSO, apenas o Intercepto e a Variável AUTO, que define a existência ou não dos altos de infração apresentaram betas maiores que zero, sendo os valores 0.0006979441 para o beta do Intercepto. Na Regressão Ridge, nenhum dos betas apresentaram valores significativos, em todos os casos os valores foram baixos. Os resultados dos betas da Regressão LASSO e da Regressão Ridge são mostrados nas tabelas 6.2, 6.3 na Seção 6 (Apêndice). Como as regressões não foram efetivas, foi realizada Análise de Componente Principais para fazer a composição das variáveis.

3.3.2 Análise de Componentes Principais

Análise de Componentes Principais (*Principal Component Analysis* - PCA) é uma técnica multivariada utilizada para analisar um grupo de dados que possuem variáveis quantitativas dependentes inter-correlacionadas e selecionar as informações importantes dos dados. A seleção das informações importantes vai gerar um novo conjunto de variáveis que são uma representação ortogonal das variáveis iniciais chamados de componentes principais Abdi e Williams (2010). Os componentes principais são combinações lineares das variáveis originais, são independente entre si e retêm a maior informação em ordem de estimação, em termos da variância total.

Para entender matematicamente como funciona a técnica, suponha uma base de dados com p variáveis e n observações que geram uma matriz $\mathbf{X}$.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

Para criar os componentes principais, são criados a partir da matriz de covariâncias dos dados (Matriz $\mathbf{S}$), baseado nas médias e nas variâncias a partir dos dados

escalados presentes na Matriz $\mathbf{Z}$, isto é, após um ajuste de média e variância dos dados mostrando na equação 3.1.

$$z_{np} = \frac{x_{np} - \mu_n}{\sigma_n} \tag{3.3}$$

$$\mathbf{Z} = \begin{pmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{np} \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} \sigma_{11}^2 & \cdots & \sigma_{1p}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \cdots & \sigma_{np}^2 \end{pmatrix}$$

Da matriz de Covariâncias $\mathbf{S}$, são gerados os autovalores (λ) por meio da equação 3.2. Para cada autovalor, é possível calcular o autovetor a, visto na equação 3.3.

$$det[Z - \lambda I] = 0 \tag{3.4}$$

$$(Z - \lambda I) * a = 0 \tag{3.5}$$

Os componentes são gerados por meio da multiplicação dos autovetores com os valores originais. Cada autovetor (a_i) há um autovalor (λ_i), o componente principal é dado por:

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$$

Cada componente possui uma contribuição C_i que é calculada pela variância Y_i dividido pela variância total, vide equação 3.4.

$$C_i = \frac{Var(Y_i)}{\sum Var(Y_i)} * 100 \tag{3.6}$$

A contribuição corresponde à importância do componente principal para o modelo, ou seja, a importância está atrelada à proporção da variância do componente perante a variância total dos componentes. A seleção dos componentes mais relevantes se

dá por meio da soma k da variância dos primeiros componentes até gerar uma proporção entre as variâncias que consiga explicar a maior parte do modelo, diminuindo a dimensão original de p para k.

Utilizando PCA na base de dados deste estudo apenas para as variáveis numéricas (Dimensão de Investimentos), o resultado gerou 16 componentes principais que são descritos de acordo com as variáveis mais correlacionadas de acordo com a tabela 3.1. Isso significa que a dimensão original diminuiu de 87 variáveis para 16 componentes principais. A Tabela 3.1 foi gerada por meio do mapa de calor (evidenciado pela Figura 3.1) da correlação entre as variáveis e os componentes. Na Seção 6 (Apêndice), a tabela 6.1 apresenta um dicionário com as variáveis utilizadas.

Componente Principal	**Variáveis mais Correlacionadas - Positivas**	**Variáveis mais Correlacionadas - Negativas**	**Nome**
PC1	retCOTASDEFUNDO1, retCOTASDEFUNDO2, retCOTASDEFUNDO3, retCOTASDEFUNDO4, retDEPOSITO1, retDEPOSITO2, retDEPOSITO3 , retDEPOSITO4	retEMPRESTIMO1, retEMPRESTIMO2, retEMPRESTIMO3, retEMPRESTIMO4	Fluxo de Saída para Emprestimos
PC2	retIMOVEL1, retIMOVEL2, retIMOVEL3, retIMOVEL4	retTITULOPUBLICO1, retTITULOPUBLICO2, retTITULOPUBLICO3, retTITULOPUBLICO4, retTITULOPUBLICO5	Politica de Investimento com menor liquidez
PC3	retOPERACAOCOMPROMISSADA1, retOPERACAOCOMPROMISSADA2, retOPERACAOCOMPROMISSADA3, retOPERACAOCOMPROMISSADA4, retOPERACAOCOMPROMISSADA5, retTITULOPRIVADO1, retTITULOPRIVADO2, retTITULOPRIVADO3, retTITULOPRIVADO4, retTITULOPRIVADO5, retCOTASDEFUNSO5	retTITULOPUBLICO1, retTITULOPUBLICO2, retTITULOPUBLICO3, retTITULOPUBLICO4, retTITULOPUBLICO5	Política de Investimento de maior risco
PC4	retACOES1, retACOES2, retACOES3, retACOES4,retACOES5, retTITULOPRIVADO1, retTITULOPRIVADO2, retTITULOPRIVADO3, retTITULOPRIVADO4, retTITULOPRIVADO5	retOPERACAOCOMPRIMISSADA1, retOPERACAOCOMPRIMISSADA2, retOPERACAOCOMPRIMISSADA3, retOPERACAOCOMPRIMISSADA4, retOPERACAOCOMPRIMISSADA5	Política de Investimento de maior risco
PC5	retVL_PAGAR_VLRECEBER1, retVL_PAGAR_VLRECEBER2, retVL_PAGAR_VLRECEBER3, retVL_PAGAR_VLRECEBER4, retVL_PAGAR_VLRECEBER5		Controle de Saída e Entrada
PC6	retTITULOPRIVADO1, retTITULOPRIVADO2, retTITULOPRIVADO3, retT5TULOPRIVADO4, retTITULOPRIVADO5, retTITULOPUBLICO1, retTITULOPUBLICO2, retTITULOPUBLICO3	retACOES1, retACOES2, retACOES3, retFINANCIAMENTO_IMOBILIARIO1, retFINANCIAMENTO_IMOBILIARIO2, retFINANCIAMENTO_IMOBILIARIO3, retFINANCIAMENTO_IMOBILIARIO4	Títulos do Mercado Financeiro
PC7	retCOTASDEFUNDO5	retACOES1, retACOES2, retACOES3, retACOES4, retACOES5	Política de Investimento consevadora
PC8	retFINANCIAMENTO_IMOBILIARIO1, retFINANCIAMENTO_IMOBILIARIO2, retFINANCIAMENTO_IMOBILIARIO3, retFINANCIAMENTO_IMOBILIARIO4, retFINANCIAMENTO_IMOBILIARIO5	retDEPOSITO5, retCONTASDEFUNDO5	Crédito Imobiliário
PC9	retDEPOSITO5	retVL_PAGAR_VLRECEBER5	Receita de Investimentos de Longo Prazo
PC10	retTITULOPRIVADO4, retTITULOPRIVADO5, retACOES4, retACOES5	retDIREITO_CREDITORIO1, retDIREITO_CREDITORIO2, retDIREITO_CREDITORIO3, retDIREITO_CREDITORIO4	Investimentos de Longo Prazo

PC11	retTITULOPRIVADO1, rEtTITULOPRIVADO2	retACOES5	Política de Investimento de maior risco
PC12	retEMPRESTIMO5	TOTAL, IDADEFUNDO	Pagamento de Empréstimos
PC13	retEMPRESTIMO5, retDIREITO_CREDITORIO4, retDIREITO_CREDITORIO5	retDIREITO_CREDITORIO1, retDIREITO_CREDITORIO2, retDIREITO_CREDITORIO3, retTITULOPRIVADO5	
PC14	retFINANCIAMENTO_IMOBILIARIO2, retFINANCIAMENTO_IMOBILIARIO3, retEMPRESTIMO5	retFINANCEAMENTO_IMOBILIARIO1	Retorno em Finanaciamento Imobiliário
PC15	retDIREITO_CREDITORIO4, retDIREITO_CREDITORIO5	TOTAL, retEMPRESTIMO5	Receita em Direito Creditório
PC16	TOTAL, retACOES1, retACOES2, retTITULO_PRIVADO5, retDIREITO_CREDITORIO5	PERIODO_PONDERADO	Total de Investimentos e Receita de Investimentos de Longo Prazo

Tabela 3.1 – Interpretabilidade dos Componentes Principais

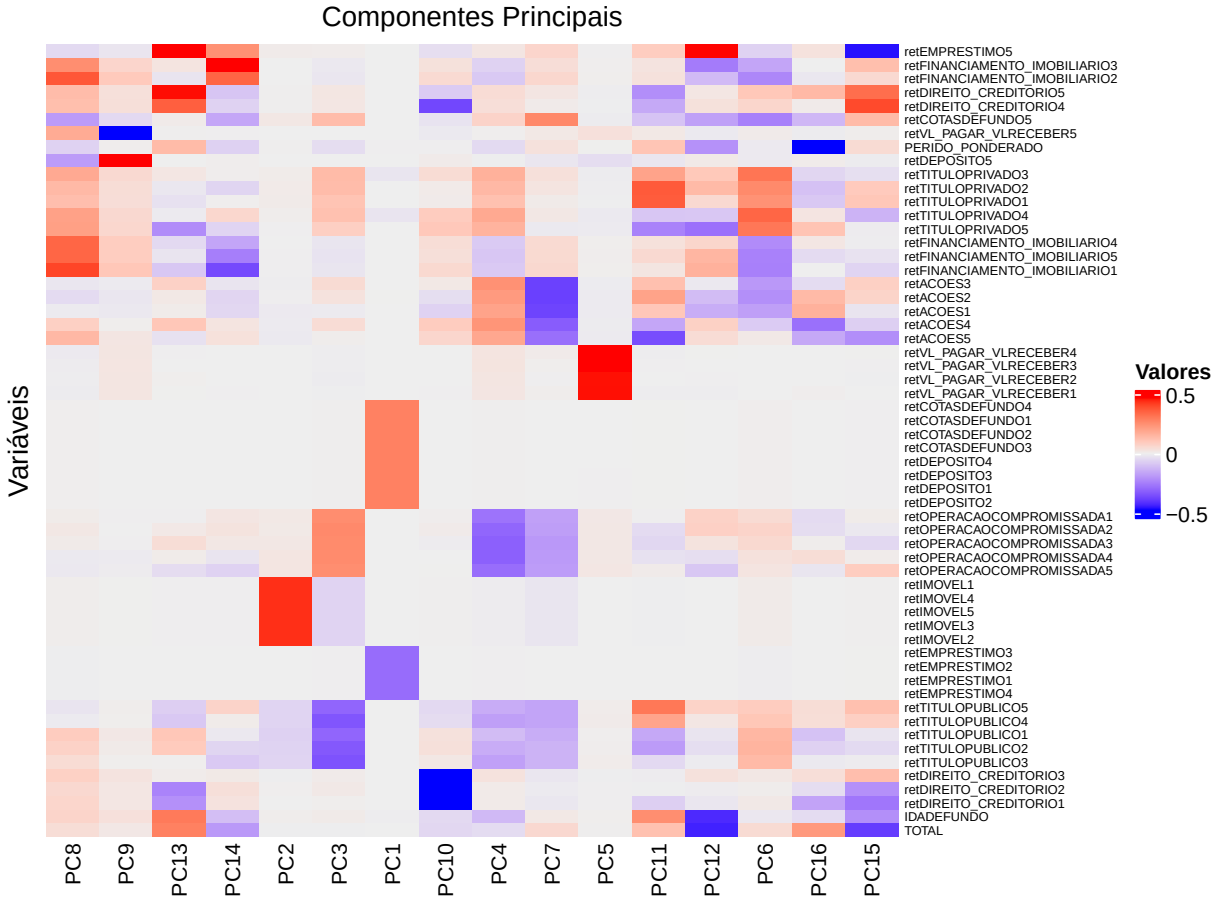

Figura 3.1 – Mapa de Calor - PCA

O Mapa de Calor 3.1 representa a relação entre as Variáveis e os Componentes principais gerados pelo PCA. No gráfico, quanto mais forte a cor, maior é a representação daquela variável dentro do Componente Principal, seja de forma positiva ou negativa. Por exemplo, no Componente Principal 1 (PC1) a variável que possui maior representação são retCOTASDEFUNDO e retDEPOSITO (Retorno em Cotas

de Fundo e Depósito defasados de 1 a 5 anos).

Os componentes são formados por correlações entre as variáveis e são indicados pelos nomes na quarta coluna da Tabela 3.1 que são as interpretações das relações das variáveis. Para não utilizar somente os componentes que descrevem os Investimentos ainda considerar todas as dimensões, foi decidido que, para uso neste estudo, foi feito a junção dos 16 componentes (relacionados a Investimentos) e das *dummies* da Dimensão de Dirigentes, com as *dummies* relacionadas a partido político, e da Dimensão do Fundo de Pensão, com as *dummies* referentes aos Patrocinadores. Para não utilizar somente os componentes que descrevem os Investimentos e ainda considerar todas as dimensões da Base de Dados, foi decidido a junção dos 16 componentes (relacionados a Investimentos) e das *dummies* da Dimensão de Dirigentes, com as *dummies* relacionadas a partido político, e da Dimensão do Fundo de Pensão do Fundo e a idade do Fundo em meses.

3.4 Modelos de Classificação Supervisionados

Após a consolidação da Base de Dados que foi utilizada, nesta Seção apresentamos os modelos que foram utilizados para prever as fraudes. Foram testados ao todo 5 Modelos de Aprendizado de Máquina que são: Regressão Logística, *Random Forest*, *LightGBM*, Máquina de Suporte Vetorial e Redes Neurais. Dentre eles, o modelo *baseline* que foi utilizado neste estudo o qual é a Regressão Logística.

Os Modelos Supervisionados necessitam de um ”exemplo“ para realizar as análises e tirar conclusões acerca do conjunto de dados. Este tipo de modelo utiliza lógica de aprendizado indutivo em que precisa de um classificador pré-estabelecido para fazer a distribuição dos dados. O classificador é construído com uma base de treinamento, que consiste em amostra do conjunto de dados, e posteriormente os dados são validados com outra amostra para verificar a acurácia dos resultados. Nas análises aqui abordadas, para a base de treinamento foi utilizado 70% da base de dados e 30% para validação.

3.4.1 Regressão Logística

Na Regressão Logística, a estimação da probabilidade de ocorrência do modelos é feita por meio da seguinte fórmula de uma função logística:

$$\sigma(x) = \frac{1}{1+e^{x}} \tag{3.7}$$

A função logística vai gerar as probabilidades de ocorrência de um fenômeno por meio de uma transformação linear $\mathbf{X}\hat{\beta}$, em que $\mathbf{X}$ é a matriz com dos valores de x e $\hat{\beta}$ são os coeficientes, que são estimados por método de máxima verossimilhança. Então, para Regressão Logística, a equação 3.8 encontra-se abaixo e, em que $\mathbf{X}$ é a matriz dos dados de entrada β são os coeficientes e as previsões são gerados por $\hat{y} = \sigma(X\hat{\beta})$.

$$y = \sigma(X\beta + \varepsilon) \tag{3.8}$$

O modelo de Regressão Logística deste estudo foi estimada por meio da função *glm()* do pacote *stats* do R, para melhorar a performance do modelo também foi utilizado *stepwise*, que consiste em um mecanismo iterativo de testa o melhor conjunto de variáveis preditoras. A ideia por trás do *stepwise* é retirar ou adicionar as variáveis preditoras afim de encontrar aquelas que trazem melhores resultados para os modelos, com menores erros. Após o *stepwise*, as variáveis que permaneceram no modelo foram: Fluxo de Saída para Empréstimos (PC1), Política de Investimento mais líquida (PC2), Política de Investimento de maior risco (PC3), Política de Investimento de maior risco (PC4), Receitas em Empréstimos e Direito Creditório (PC13), Retorno em Financiamento Imobiliário (PC14), Receita em Direito Creditório (PC15), lagAVANTE1, lagCIDADANIA1, lagMDB1, lagPSDB1, lagPL1, lagPSC1, lagPSL1,lagPTB1, Instituidor, Privada, Pública Estadual, Pública Federal

3.4.2 Random Forest

Como já citado no Referencial Teórico, o Random Forest é um modelo que se baseia em um conjunto árvores de decisão, em que cada árvore “cresce” a partir de uma combinação de variáveis aleatórias. Suponhamos um vetor aleatório de entrada $X = (X_1, X_2, ..., X_p)^n$ e Y como a variável resposta real, podemos afirmar que há

uma distribuição desconhecida que modela $P_{XY}(X,Y)$.O objetivo do modelo é encontrar uma função $f(X)$ que consiga prever Y e esta função é definida pela minimização do valor médio de uma função de perda entre aquilo que é predito e a realidade $E_{XY}(L(Y,f(X)))$.$L(Y,f(X))$ é uma medida para o quão perto $f(X)$ está de Y e penaliza quando o valor é muito distante (CUTLER; CUTLER; STEVENS, 2012). Para classificação, é definido como

$$L(Y,f(X)) = I(Y \neq f(X)) = \begin{cases} 0 & \text{if} Y = F(X) \\ 1 & \text{caso contrário} \end{cases} \tag{3.9}$$

Em problemas de classificação, para minimizar $E_{XY}(L(Y,f(X)))$ seguindo a *zero-one loss* se dá por meio de $\{h(x,\Theta_k), k = 1,...\}$

$$f(x) = \arg\max P(Y = y|X = x) \tag{3.10}$$

A funções preditora $f(x)$ são construídas por árvores de decisão $\{h(x,\Theta_k), k = 1,...\}$. No caso da classificação, a função preditora do conjunto é definida pela classe mais frequentemente predita, isto é, pelos votos no resultado final de casa árvore.

$$f(x) = \arg\max \sum_{j=1}^{J} I(y = h_j(x)) \tag{3.11}$$

Para selecionar os melhores parâmetros do modelos, foi utilizado *cross-validation* para os parâmetros *ntree*, que é o número de árvores; *mtry*, que é o número de variáveis amostradas em cada divisão; *maxnodes*, que é o número máximo de árvores de nós terminais da floresta; *nodesize*, que é o tamanho mínimo dos nós terminais. O *cross-validation* usou a Acurácia como métrica para selecionar os parâmetros. A modelagem do *Random Forest* foi realizada no *software R* com o pacote *randomForest*.

3.4.3 Máquina de Suporte Vetorial

Para fazer as previsões nos algoritmos de Máquina de Suporte Vetorial (SVM) é necessária uma função de decisão para classificação com objetivo de maximizar a margem entre as classes. O modelo mais simples de SVM seria uma função de classificação binária representada por $f(\mathbf{x}) = sinal(\mathbf{w}^T\mathbf{x} - \gamma)$. O algoritmo do SVM utilizará um vetor de dados $\mathbf{x}$ de dimensão p x 1, em que p é quantidade de variáveis, e $\mathbf{w}$ que

é um vetor de parâmetros de mesma dimensão e γ é o termo de viés. Com objetivo de fazer a estimação da máxima margem entre as classificações, é necessário definir $\mathbf{w}$ e γ que façam a separação entre as classes. Como é possível observar com a Figura 3.2

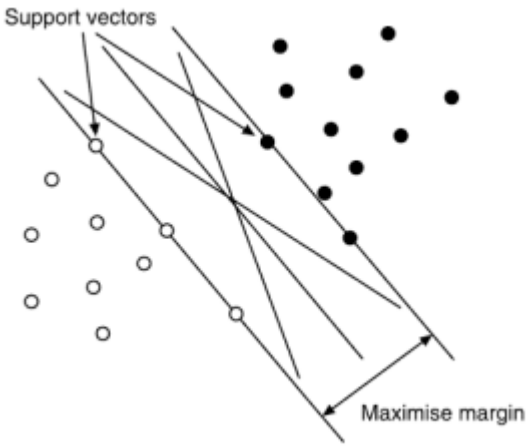

Figura 3.2 – Separador de Máxima Margem
Fonte: Soman, Loganathan e Ajay (2009)

A máxima margem é representada por ζ na equação 3.13 definida por um problema de separação linear, em que o *input* é a matriz de dados $\mathbf{X}$ com dimensões n x p em que o objetivo é fazer previsão de uma variável y de dimensão n x 1 que contém as classes 0 e 1 (ALBUQUERQUE, 2014). Para um caso linear como na Figura 3.2, o SVM consegue ser resolvido por meio de um Problema de Programação Linear:

$$\begin{aligned} &\text{Maximizar} && \zeta = \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &\text{Sujeito a} && \mathbf{D}(\mathbf{X}\mathbf{w} - \gamma\mathbf{1} \geq 1 \\ &\text{Com} && \mathbf{w} \in \mathbb{R}^{\text{l}};\quad \gamma \in \mathbb{R} \end{aligned} \tag{3.12}$$

Em que $\mathbf{1}$ é um vetor unitário de dimensão n x 1 e $\mathbf{D} = diag(y)$. A equação 3.13 é resolver um problema linear, para solucionar o problema que também lide com não-linearidade, ao invés de considerar o vetor de dados $\mathbf{x}$, considerar $\phi(\mathbf{x})$, de modo que $\phi(\mathbf{x}) \in \mathbb{R}^q$ e $q > p$, ou seja, os dados serão mapeados para dimensões maiores

($\mathbf{R^p} \rightarrow \mathbf{R^q}$) por meio de funções *kernel* (ALBUQUERQUE, 2014).

$$\begin{aligned} &\text{Minimizar} \quad \zeta* = \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &\text{Sujeito a} \quad \mathbf{D}(\Phi\mathbf{w} - \gamma\mathbf{1}) \geq \mathbf{1}, \text{para } \mathbf{w} \in R^q, \gamma \in R, \\ &\text{Com} \quad \Phi = \begin{matrix} \phi(\mathbf{x}_1)^T \\ \phi(\mathbf{x}_2)^T \\ \vdots \\ \phi(\mathbf{x}_n)^T \end{matrix} \end{aligned} \tag{3.13}$$

Para selecionar os parâmetros do SVM, também foi utilizado *cross-validation* utilizando a Acurácia como métrica para definir o C, isto é o parâmetro Custo que determina possíveis erros de classificação, e o $sigma$ por meio do *kernel svmRadial* do pacote *kernlab*. Para o modelo, foi utilizado o *kernel* Gaussiano. No modelo de SVM, a base manipulada foi escalada com objetivo de normalizar a base de dados, com isso algumas variáveis foram excluídas da análise por não apresentarem variância nula. Neste caso, as variáveis diminuíram de 43 para 34, sendo retirado da análise as variáveis:lagPMN1, lagPMB1, lagPRTN1, lagPROS1, lagPSD1, lagPSEL1, lagPSTU1, lagPSB1, lagPV1. Todas elas da Dimensão de Partidos Políticos. A modelagem do SVM foi realizada no *software R* por meio do pacote *caret, kernlab* e *e1071*.

3.4.4 LightGBM

O LightGBM é um tipo de estrutura de modelo *Gradient Boosting* que, assim como Random Forest, também utiliza lógica de aprendizado em árvore. A diferença principal entre o LightGBM e *Gradient Boosting* é que o LightGBM cresce verticalmente, isto é, *leaf-wise*, enquanto outros modelos crescem *level-wise*, isto faz com que o modelo seja mais rápido e ocupe menos memória. O LightGBM utiliza duas técnicas para aprimorar o *Gradient Boosting*: Amostragem unilateral baseado em Gradiente (*Gradient-based One-Side Sampling* - GOSS) e Pacote de Recursos Exclusivos (*Exclusive Feature Bundling* - EFB) (KE et al., 2017).

A Amostragem unilateral baseado em Gradiente (GOSS) é utilizada observando que dentro do conjunto há instâncias de dados com gradientes diferentes e que possuem desempenhos diferentes no ganho de informações. Isto significa que a instância com gradientes maiores, terão maior contribuição no ganho de informação, logo a amostragem focando nestas instância é benéfica para uma maior acurácia. GOSS

é uma técnica que vai focar nas instâncias de dados com gradientes maiores e eliminando aleatoriamente as instâncias com gradientes menores (KE et al., 2017).

O Pacote de Recursos Exclusivos é uma técnica utilizada quando há um grande número de *features* no conjunto de dados, porém o espaço de *features* é escasso. Muitas *features* dentro do espaço acabam sendo exclusivas, isto é, em poucas situações recebem valores diferentes de zero simultaneamente. O Pacote de Recursos Exclusivos faz parte de um algoritmo eficiente que seleciona as *features* exclusivos em uma única *feature* e assim auxiliar com o uso de menos memória no processamento (KE et al., 2017).

Para selecionar os parâmetros do modelo neste estudo, foi utilizado *cross-validation* para os seguintes parâmetros: *num_leaves*, isto é, o número total de folhas da árvore; *learning_rate*, que determina o impacto de cada árvore no resultado final; *min_data_in_leaf*, que estipula o mínimo de dados em uma folha, este parâmetro é utilizado para evitar *under-fitting*; *max_depth*, que limite a profundidade da árvore. Como métrica para selecionar os parâmetro, foi utilizado *AUC*. No LightGBM, assim como no SVM, também foi utilizada a base escalada para normalização dos dados e neste modelo, diferentemente dos outros, a modelagem foi realizada no *software Python 3.7* com o pacote *lightgbm*.

3.4.5 Redes Neurais

O modelo de Redes Neurais consiste em um conjunto de neurônios de entrada que são repassados por mais camadas ocultas de neurônios que reprocessam dados até uma camada de neurônios de saída que geram os resultados. Isso significa que a Rede Neural é composta por diferentes tipos de arquitetura em que há uma camada de neurônios de entrada, camadas ocultas de neurônios de processamento, dependendo da arquitetura escolhida e uma camada de neurônios de saída. A equação 3.14 representa a conexão, chamada de peso, entre os neurônios em que w_i é a saída de um neurônio i dentro de uma camada oculta (WANG, 2003).

$$w_i = \sigma\left(\sum_{j=1}^{N} V_{ij}x_i + T_i^{hid}\right) \tag{3.14}$$

Na equação 3.14 σ representa uma função de ativação que introduz não linearidade à rede neural, N o número de neurônios de entrada, V_{ij} os pesos, x_i os dados de

entrada, T_i^{hid} os limites dos neurônios ocultos (WANG, 2003).

Para o modelo de Redes Neurais, foi utilizado o pacote *keras* do *Python* com três camadas ocultas e utilizando a função de ativação *tanh*, que se mostrou a função com melhores resultados em comparação com a função *relu*. Na camada de *output* foi utilizado a função de ativação *sigmoid* que retorna uma saída de apenas uma uma dimensão. Como função perda, foi usado *binary_crossentropy*, como otimizador o *adam* e como métrica a *acurácia*. No treinamento do modelo, foram feitas 100 *epochs*.

4 RESULTADOS

4.1 Simulação

Antes de efetivamente aplicar os modelos, foi feito uma simulação de bases de dados em vários níveis de balanceamento para verificar o comportamento da métricas de avaliação dos modelos em cada um dos níveis. Além de diferentes níveis de balanceamento, foram considerados vários tamanhos de amostras para testar como as métricas de avaliação se comportam entre os modelos com diferentes níveis de desbalanceamento. As classificações dos dados nem sempre são bem discriminadas semanticamente, isso significa que as classes dos conjunto de dados são semelhantes mostrando pouca separabilidade entre categorias, o que pode dificultar ainda mais a previsão (ARINO; KIKUTA, 2018; RAJ; MAGG; WERMTER, 2016). A similaridade, ou falta de separabilidade, entre as classes pode prejudicar a classificação porque o aprendizado do modelo fica enviesado pelas classes majoritárias (RAJ; MAGG; WERMTER, 2016).

A simulação de dados foi feita utilizando a função *twoClasssim()* do pacote *caret* no Software Livre *RStudio*, ajustando os parâmetros para cada nível de balanceamento. Foram testadas bases com 50%, 40%, 30%, 20%, 10%, 5% de balanceamento e com amostras de 1000 observações.

De forma inicial, foi aplicado Regressão Logística, *Random Forest* e Máquina de Suporte Vetorial para verificar o desempenho dos modelos para dados desbalanceados. Foram usadas as métricas: Acurácia, Sensibilidade ou *Recall*, Especificidade e *Area Under the Curve* (AUC). Todas elas são utilizadas para Modelos Supervisionados, isto é, nas Abordagens Reativas. As métricas são decorrentes da Matriz de Confusão, ou Matriz de Erro, que consiste em uma tabela de visualização do desempenho do modelo, como pode ser visto no exemplo da tabela 4.1.

A Acurácia é a métrica mais utilizada para estimar o desempenho de Modelos de Classificação. A equação 4.1 representa a fórmula da Acurácia e, por mais que

		Classe Predita	
		1	0
Classe Real	1	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	0	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Tabela 4.1 – Matriz de Confusão

seja uma medida muito utilizada da literatura da área, em dados desbalanceados a métrica não se adéqua por possuir um viés de preferência às classes majoritárias, desconsiderando o impacto da das classes minoritárias do modelo (BRANCO; TORGO; RIBEIRO, 2015).

$$\text{Acurácia} = \frac{VP+VN}{FN+FP+VP+VN} \tag{4.1}$$

A Sensibilidade ou *Recall*, representada pela Equação 4.2, e a Especificidade, representada pela Equação 4.3, são métricas complementares. Sensibilidade de refere aos Verdadeiros Positivos do modelo e mede o desempenho da classe positiva. A Especificidade se refere aos Verdadeiros Negativos e mede o desempenho da classe negativa. Para modelos desbalanceados, a métrica mais interessante seria a Sensibilidade para verificar a performance dos positivos da análise.

$$\text{Sensibilidade ou } \textit{Recall} = \frac{VP}{VP+FN} \tag{4.2}$$

$$\text{Especificidade} = \frac{VN}{FP+VN} \tag{4.3}$$

A AUC também é derivada da Matriz de Confusão e da Curva ROC. A Curva ROC ou *Receiver Operating Characteristic Curve* é uma representação gráfica utilizando a Taxa de Verdadeiros Positivos (uma representação da Sensibilidade) e a Taxa da Falsos Positivos (1 - Especificidade). A AUC consiste na área sobre a curva ROC (*Area Under the Curve*). A curva ROC possui uma forma exponencial e pode ser ajustada pela estimativa iterativa de Máxima Verossimilhança. Este ajuste que auxilia na estimativa da AUC. Para todas estas métricas, quanto mais próximo o valor estiver de 1, melhor o desempenho do modelo.

Na simulação realizada com Regressão Logística, ficou perceptível que os comportamentos das métricas são diferentes para dados que possuem um certo nível de balanceamento, por exemplo 50%, 40% ou 30% em uma das classes, que para bases

mais desbalanceadas, por exemplo 20%, 10% e 5%. As métricas que mais se relacionaram foi o *Recall* e a Especificidade. A medida que a proporção entre as classes se diferencia muito o comportamento do *Recall* e da Especificidade se torna oposto, ou seja, comforme o desbalanceamento aumenta a Especificidade vai aumentando e o *Recall* vai diminuindo. Isso acontece porque Especificidade está relacionado com os Verdadeiros Negativos da Matriz de Confusão, que são os valores das classes majoritárias da análise (No caso deste estudo a não-ocorrência de fraude). Já o *Recall* diminui porque a métrica calcula o desempenho de Verdadeiros Positivos da Matriz de Confusão. Em dados desbalanceados, a Regressão Logística generaliza a classificação enviesando para as classes majoritárias, logo, os Verdadeiros Positivos (no caso desde estudo, a fraude) são baixos e, por isso, o *Recall* diminui.

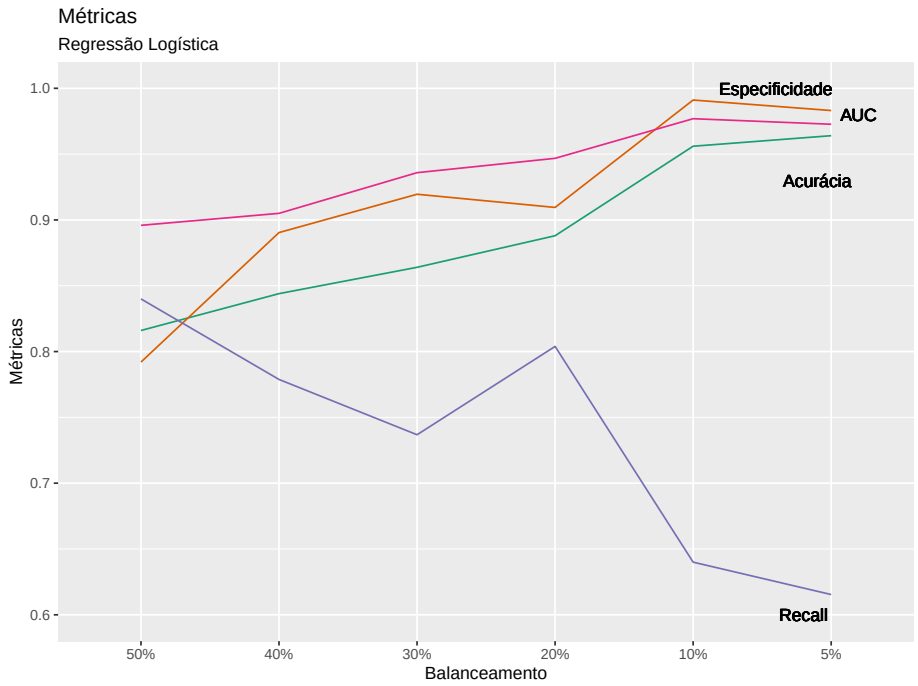

Figura 4.1 – Métricas - Regressão Logística

	50%	40%	30%	20%	10%	5%
Acurácia	0,82	0,84	0,86	0,89	0,96	0,96
Especificidade	0,79	0,89	0,92	0,91	0,99	0,98
Recall	0,84	0,78	0,74	0,80	0,64	0,62
AUC	0,90	0,90	0,94	0,95	0,98	0,97

Tabela 4.2 – Métricas - Regressão Logística

Para os resultados do *Random Forest*, as bases extremamente desbalanceadas,

com 10% a 5%, o modelo não apresentou bons resultados em termos de *Recall*, Especificidade e AUC. O resultado da Acurácia se mostrou maior de 80% para todos os balanceamentos, o que pode ser considerado um bom resultado. Isso se dá pelos problemas, já citados anteriormente, de generalização dos modelos. A Acurácia é uma medida que relaciona os valores dos Verdadeiros Positivos e Verdadeiros Negativos com soma dos valores totais da Matriz de Confusão. Por isso, em bases desbalanceadas esta métrica acaba mascarando os resultados. O *Recall* e a Especificidade obtiveram o mesmo comportamento apresentando na Figura 4.1 do Resultado da Regressão Logística. Observando o comportamento do *Random Forest* para dados desbalanceados, para este modelo sugere-se o uso de algum tipo de ferramenta estatística para "balancear" as bases como SMOTE ou utilizar os modelos adaptados como *Balanced Random Forest* ou *Weighted Random Forest*.

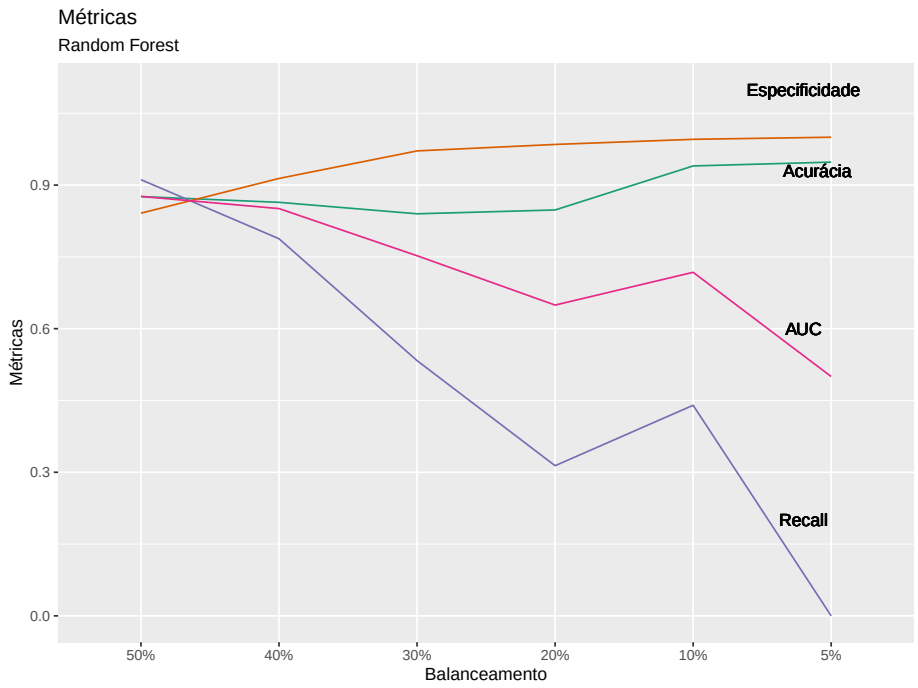

Figura 4.2 – Métricas - *Random Forest*

	50%	40%	30%	20%	10%	5%
Acurácia	0,88	0,86	0,84	0,85	0,94	0,95
Especificidade	0,84	0,91	0,97	0,98	1,00	1,00
Recall	0,91	0,79	0,53	0,31	0,44	0,00
AUC	0,88	0,85	0,75	0,65	0,72	0,50

Para Máquina de Suporte Vetorial, o modelo apresentou métricas razoáveis de

50% a 10% de balanceamento das classes. Na Tabela 4.3, as métricas mostram que a Acurácia e a Especificidade permanece estável para todos os níveis de balanceamento das bases, Entretanto, o *Recall* e a AUC tem uma queda significativa entre os balanceamento de 50% a 5%. Estes resultados mostram que o *Recall* e AUC são métricas importantes para medir o desempenho de modelos que lidam com altos níveis de desbalanceamento de bases de dados, principalmente para mostrar o desempenho dos Verdadeiros Positivos do resultado. Novamente, o comportamento da Especificidade e do *Recall* é oposto a medida que o desbalanceamento aumenta.

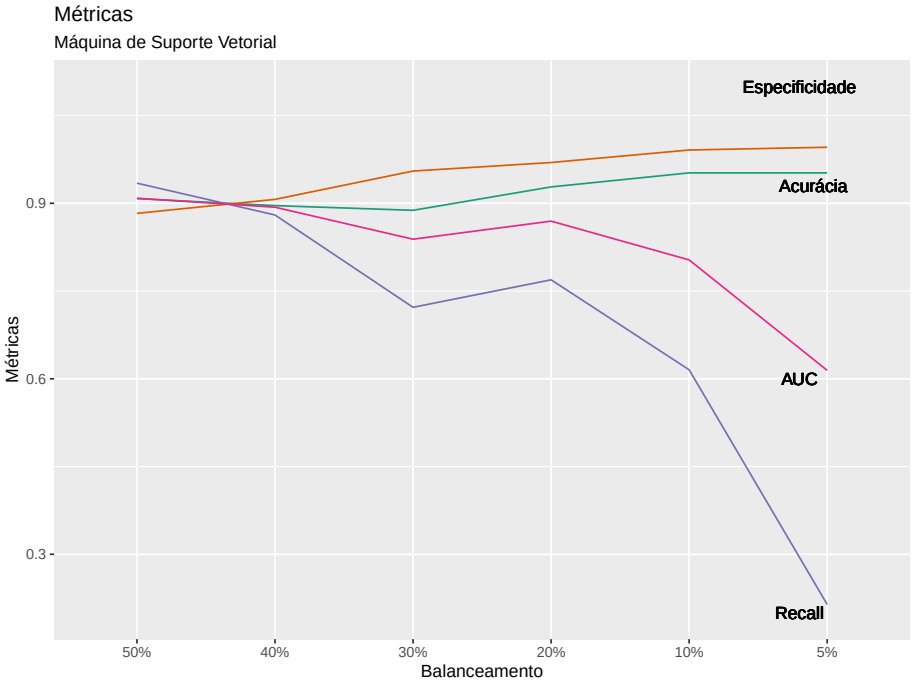

Figura 4.3 – Métricas - Máquina de Suporte Vetorial

	50%	40%	30%	20%	10%	5%
Acurácia	0,91	0,90	0,89	0,93	0,95	0,95
Especificidade	0,88	0,91	0,96	0,97	0,99	1,00
Recall	0,93	0,88	0,72	0,77	0,62	0,21
AUC	0,91	0,89	0,84	0,87	0,80	0,61

Tabela 4.3 – Métricas - Máquina de Suporte Vetorial

Com a simulação, é possível fazer constatações acerca do comportamento das métricas de acordo com o balanceamento das bases. A Acurácia, amplamente, utilizada na literatura da área, não apresenta bons resultados. A medida traz um viés

para as classes majoritárias, por ser uma métrica global que avalia tanto os Verdadeiros Positivos como os Verdadeiros Negativos. Como os Verdadeiros Negativos são muito altos, os valores “mascaram” o resultado baixo dos Verdadeiros Positivos. A Especificidade se mostra como métrica importante para medir o desempenho dos Verdadeiros Negativos do Modelo, dependendo da resposta a ser respondida pela previsão é uma informação relevante para as interpretação pretendida. O *Recall* também é uma métrica importante para mensurar a ocorrência do evento, isto é, os Verdadeiros Positivos. De maneira prática, as duas métricas englobam conceitos importantes para avaliar os modelos, cada uma de acordo com a proposta específica.

A avaliação do modelo é uma parte importante da modelagem porque envolve a verificação do desempenho dos resultados. Se considerarmos apenas a Especificidade como métrica, o que se escolhe como avaliação é a não-ocorrência do evento e se considerarmos apenas o *Recall* podemos gerar uma avaliação de resultado que se ajusta demais a ocorrência do evento e não se aqueda a outras realidades, se tornando um modelo muito específico. Logo, pela simulação ficou evidente que ambas as métricas devem ser consideradas na avaliação dos Modelos de Classificação quando são analisados dados com alto desbalanceamento entre as classes.

4.2 Técnicas de Balanceamento

Uma solução possível para problemas com bases desbalanceadas é o uso de Técnicas de Balanceamento da bases de dados parar ajustar a proporção entre as ocorrências das classes. Como observado na descrição da Base de dados na Seção 3.2, os dados apresentam um alto desbalanceamento. Com isso, entendeu-se a necessidade do uso de técnicas de reamostragem para balancear as bases, sejam técnicas de *Under-Sampling*, que fazem a reamostragem removendo da base de dados amostras que pertençam à classe majoritária, ou técnicas de *Over-Sampling*, que fazem re-amostragem aumentando o número de instâncias da classe minoritária (MOHAMMED; RAWASHDEH; ABDULLAH, 2020). Os testes para analisar as Técnicas de Balanceamento foram realizados antes da estimação dos modelos para auxiliar na tomada de decisão das Técnicas que seriam usadas para balancear a base.

Para avaliar qual seria a melhor Técnica de Balanceamento para implementar, foi realizado um teste utilizando o *toolkit open-source* do *Python* chamado *imbalanced-learn* que tem por objetivo fornecer uma variedade de métodos para balancear um

conjunto de dados que esteja desbalanceado. Os métodos implementados estão em 4 grupos principais: *Under-Sampling*, *Over-Sampling*, Combinação de *Under-Sampling* e *Over-Sampling* e Métodos de Aprendizado por Conjunto ou Métodos *Ensemble* (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017).

Para o teste, foi selecionada uma amostra da Base de Dados com variáveis dos Partidos Políticos e dos Ativos Financeiros com apenas 1 ano de defasagem. Foi considerado 1 ano de defasagem porque entende-se que os acontecimentos de decorrem em uma fraude ocorrem com uma janela de antecedência do evento. O Modelo de *benchmark* que foi utilizado foi a Regressão Logística com diferentes porcentagens de treinamento de 50%, 40%, 30%, 20% e 10%. As porcentagens diferentes de treinamento e validação foram testadas para verificar as possibilidades de resposta do modelo para Falsos Negativos ou Falsos Positivos, dentro da Matriz de Confusão. Os métodos de Balanceamento testados foram: *Under Sampling* com *Cluster Centroids* (ZHANG; ZHANG; WANG, 2010), *Condensed Nearest Neighbour* (HART, 1968), *Edited Nearest Neighbours* (WILSON, 1972) , *AllKNN* (TOMEK et al., 1976), *Instance Hardess Threshold* (SMITH; MARTINEZ; GIRAUD-CARRIER, 2014; VERDIKHA; ADJI; PERMANASARI, 2018), *Bordeline SMOTE* (HAN; WANG; MAO, 2005), *Random Oversampler* (BATISTA; PRATI; MONARD, 2004), *Random Undersampling* (MISHRA, 2017), *SMOTE*(CHAWLA et al., 2002), *Support Vector Machine SMOTE* (WANG, 2008; NGUYEN; COOPER; KAMEI, 2011), *SMOTE + ENN* (BATISTA; PRATI; MONARD, 2004) e *SMOTE + TomekLinks* (BATISTA et al., 2003).

As métricas selecionadas para avaliar o desempenho do modelo foram Acurácia, Especificidade, Sensibilidade ou *Recall* e *Geometric Mean Score*. A Acurácia é uma métrica extensamente na literatura da área para medir o desempenho de modelos de Aprendizado de Máquina, principalmente por ser uma técnica global que considera todos os Verdadeiros e Falsos Positivos e Verdadeiros e Falsos Negativos da Matriz de Confusão, como já citado na Sessão 4.1. Entretanto, para uso em bases desbalanceadas a métrica não possui um desempenho tão bom por causa da tendência de generalização das classes majoritárias.

Como nova métrica para análise, dado os resultados da Simulação na sessão 4.1, foi usado o *Geometric Mean Score* (G-Mean) . O G-Mean consiste na Média Geométrica entre a Especificidade e a Sensibilidade. A Especificidade mede o desempenho dos Verdadeiros Negativos da Previsão, enquanto a Sensibilidade mede o desempenho dos Verdadeiros Positivos da Previsão, esta última medida é importante para avaliar as previsões de bases desbalanceadas, tendo em vista que como a ocorrência do

fenômeno é minoritária na base de dados, e por isso é interessante uma medida que consiga avaliar como estas previsões dos dados minoritários desempenharam na previsão.

Observando a a importância da Sensibilidade, o G-Mean se torna uma métrica muito relevante para verificar o equilíbrio entre as duas medidas que conseguem avaliar o desempenho dos Verdadeiros Positivos e dos Verdadeiros Negativos do modelo. Devido a importância entre o equilíbrio das duas métricas, para selecionar o método de balanceamento que será utilizado neste estudo, usaremos o método que obteve o maior valor de G-Mean para o teste com Regressão Logística. A equação do G-Mean encontra-se a seguir:

$$\text{G-Mean} = \sqrt{\text{Sensibilidade} * \text{Especificidade}} \tag{4.4}$$

Um estudo realizado por Luque et al. (2019) mostrou que a métrica *Geometric Mean Score* e *Informedness Bookmaker* são as medidas que apresentam menor viés para medir desempenho em modelos de classificação para dados desbalanceados, em que o foco é o sucesso da classificação dos resultados. As métricas de Sensibilidade e Especificidade também são medidas que não possuem viés por desequilíbrio, entretanto, são métricas de desempenho unidimensionais porque consideram apenas os resultados da classe negativa ou da classe positiva e não em ambas as classes. Observando isso, o uso da *Geometric Mean Score* soluciona o problema da unidimensionalidade para medir o sucesso da classificação e mantém a vantagem de não possuir viés de desequilíbrio como as Métricas de Acurácia e Precisão (LUQUE et al., 2019).

Nos Gráficos 4.4 e 4.5, estão o resultado do teste realizado com as Técnicas de Balanceamento. A abcissa compreende as Porcentagens da Base de dados que foi Utilizada para Validação do modelo de Regressão Logística e a ordenada representa o Valor do G-Mean Acumulado no Gráfico 4.4 e o Valor Absoluto da Métrica no Gráfico 4.5. Como os resultados é possível observar que o Comportamento do G-Mean não se modifica drasticamente com o aumento da base e Validação, isto fica claro pelo formato das áreas dos Modelos nos Gráfico 4.4. Caso houvesse uma diferença drástica entre as porcentagens de Base de Validação, as áreas deveriam apresentar formatos mais caóticos que os formatos apresentados.

Outro resultado que pode ser interpretado baseado nos Gráficos gerados é que, para a base de dados utilizada, as Técnicas de Balanceamento que obtiveram maiores

G-Mean foram *Random Undersampling*, para 50% e 10% da Base usada para Validação; SMOTETomek, para 30% da Base de Validação; e SMOTE também pata 30% da Base para Validação. Tendo em vista que não houve grande diferença nos resultados para os diferentes tamanhos de base de Validação, pode-se considerar que estas três técnicas seriam as mais adequados parar serem utilizadas para balancear as bases. As três técnicas serão mais detalhadas nas seções a seguir.

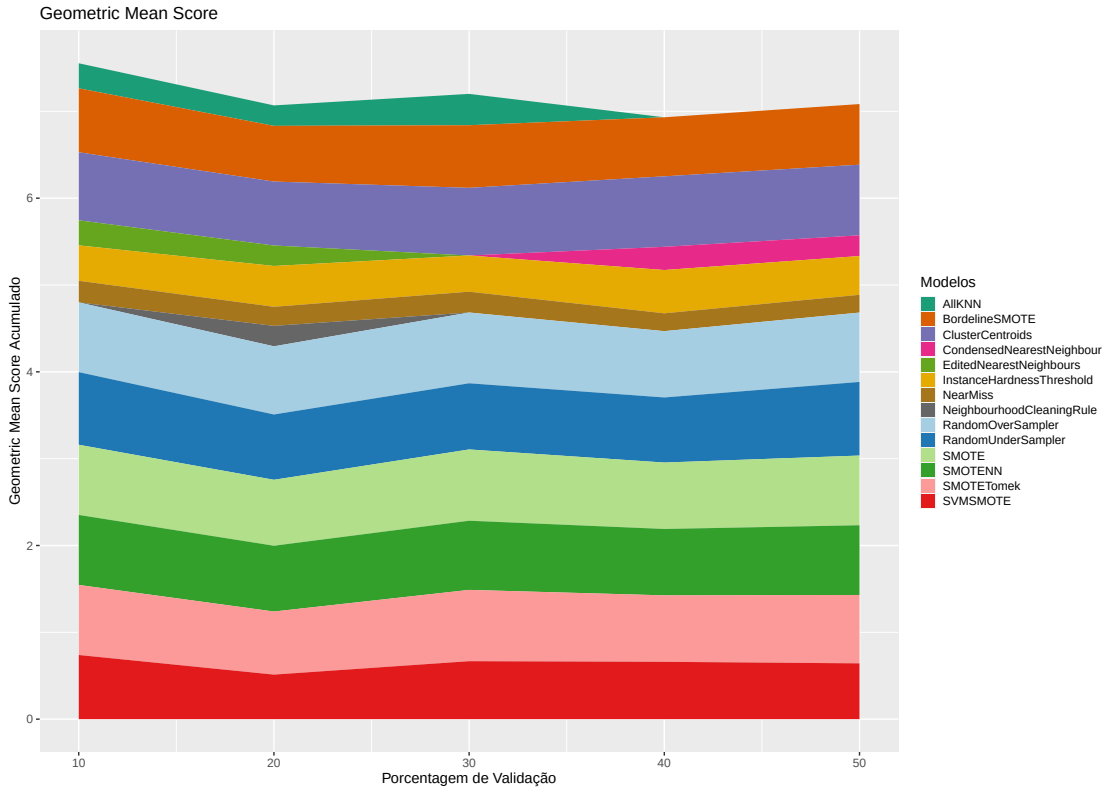

Figura 4.4 – Gráfico de Comportamento do *Geometric Mean Score* Acumulado para Diferentes tipos de Validação

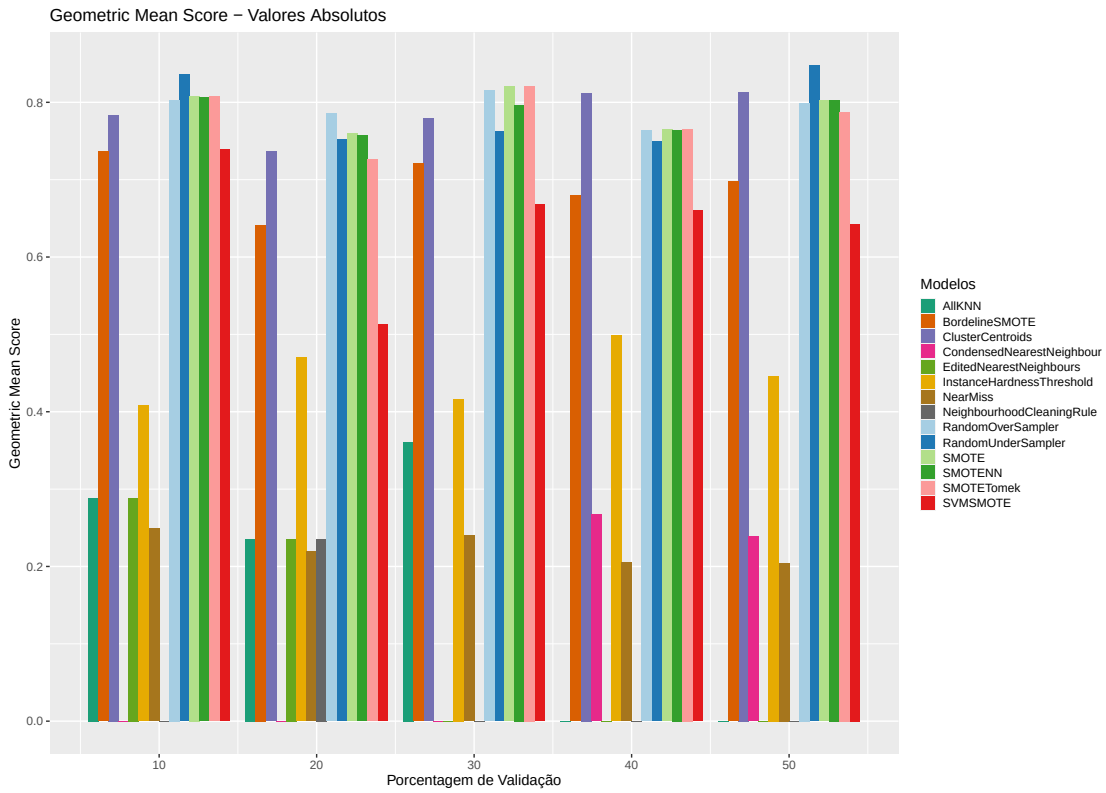

Figura 4.5 – Gráfico de Comportamento do *Geometric Mean Score* Absoluto para Diferentes tipos de Validação

4.2.1 *Random Undersampling*

As técnicas de re-amostragem são utilizadas para balancear as bases seguem basicamente uma das duas lógicas a seguir: diminuir a incidência da classe majoritária (*undersampling*) ou aumentar a incidência da classe minoritária (*oversampling*). O *Random Undersampling* consiste em uma remoção aleatória dos dados com classe majoritária ou de uma sub amostra da classe majoritária, mantendo a classe minoritária com a população completa (HASANIN; KHOSHGOFTAAR, 2018). Isso significa que a técnica retira informações da base de dados para equilibrar a diferença entre as classes, diminuindo a população da base de dados. Uma das desvantagens do uso desta técnica é que o método pode descartar dados com potencial importância para explicar os fenômenos (BATISTA; PRATI; MONARD, 2004). Para este estudo, a técnica selecionou aleatoriamente da 99.5% de não-ocorrência de fraude e gerou uma base de dados com 114 observações para a base de treino do modelo e com balanceamento de 50% para ocorrência da incidência de fraude e não-ocorrência. O *Random Undersampling* acarreta em uma diminuição considerável da base de dados, por mais que seja uma técnica amplamente usada dentro do contexto de Dados Desbalanceados.

4.2.2 SMOTE - *Synthetic Minority Over-sampling Tchnique*

A premissa básica da técnica SMOTE é a inserção de dados sintéticos no conjunto de dados original para para aumentar as classes minoritárias (KÖKNAR-TEZEL; LATECKI, 2009). Os dados sintéticos são gerados a partir das características das classes minoritárias e do vizinho k-mais próximo (HANIFAH; WIJAYANTO; KURNIA, 2015). O procedimento ocorre da seguinte forma: é calculada a diferença entre o vetor de características e seu vizinho mais próximo e esse valor é multiplicado por um número aleatório entre 0 e 1, após isso, o resultado da multiplicação é adicionado ao vetor de características considerado (CHAWLA et al., 2002).

O cálculo mostra que operações são feitas no “espaço de características” e não no “espaço de dados” e a abordagem força a aleatoriedade dos dados sintéticos evitando que se tornem enviesados (CHAWLA et al., 2002). Na equação a seguir é possível verificar um exemplo de como o SMOTE é aplicado:

$$(X,Y) = (x_1,y_1) - \sigma * (x_2,y_2) \tag{4.5}$$

Em que X,Y são os vetores de características gerados por meio de (x_1,y_1). Com a aplicação do SMOTE, foi gerado uma base da dados de 3599 observações para a base de treinamento com balanceamento de 65% para não ocorrência de indício de fraude e 35% para ocorrência de indício de fraude.

4.2.3 SMOTE Tomek Links

O SMOTE Tomek Links é junção de duas técnicas: SMOTE e Tomek Links. A técnica de Tomek Links consiste em um método de limpeza para retirar dados que sejam ruídos e *outliers*. O algoritmo possui uma lógica semelhante a *k-Nearest Neighbor* seleciona pares de instâncias de classes diferentes, no caso deste estudo um instância da classe majoritária e da classe minoritária, e verifica a distância entre o par selecionado e remove um dos pares ou ambos para limpar o conjunto de treinamento. Um ”Tomek link“ é definido da seguinte forma: dado x e y sendo instâncias de classes diferentes, e $d(x,y)$ como a distância entre x e y, em caso de não haver uma instância z em que $d(x,z) < d(x,y)$ ou $d(y,z) < d(x,y)$ então x e y são Tomek Link e uma das instâncias são ruídos ou estão na ”linha de fronteira“ (BATISTA et al., 2003). No SMOTE Tomek Links, primeiramente é aplicado o SMOTE na base de dados de treinamento para balancear

a distribuição e após isso é utilizado o Tomek Links para limpar a base de dados ruidosos. Com o SMOTE Tomek, foi gerado uma base de treinamento de 1239 observações com um balanceamento 48% para não ocorrência e 52% para ocorrência do indício de fraude.

Para todos os modelos, na base de teste que foi utilizada não passou por nenhuma das técnicas de balanceamento para que a validação não fosse impactada. A base de Teste foi gerada com 30% dos dados da base original e possui 5198 observações. Para alguns modelos (no caso, *Random Forest*), foi realizado um filtro na base de teste para utilizar somente as informações de Fundos de Pensão que fossem Públicos com objetivo de diminuir o tamanho da base de teste, tendo em vista que devido ao uso da técnicas de balanceamento a volumetria dos dados da base treinamento original diminuiu consideravelmente em comparação à volumetria da base de dados de teste.

4.3 Previsão dos Modelos Supervisionados

O objetivo deste estudo é aplicar Modelos de Aprendizado de Máquina Supervisionados e Técnicas de Balanceamento de dados para Prever Indício de Fraude em Fundos de Pensão. Como fraudes são eventos raros, foi necessário o uso de certos tratamentos para realizar a modelagem. Estas técnicas foram descritas nas sessões anteriores, como as tentativas da Regressão LASSO e Regressão Ridge e, por fim, foi selecionada a Análise de Componentes Principais como método de tratamento dos dados para reduzir a grande quantidade de *features* e formar um base de dados com representação das informações mais importantes.

Para definir qual o Modelo Supervisionado é o melhor para previsão de dados desbalanceados dentro do contexto de fraude, usou-se o *Geometric Mean Score* (G-Mean) como métrica de análise para avaliar o desempenho do método, por ser a métrica que avalia o equilíbrio entre a Sensibilidade e a Especificidade, duas medidas que conseguem avaliar o desempenho dos Verdadeiros Positivos e dos Falsos Negativos do modelo. Para direcionar a decisão de qual será o melhor modelo baseado na métrica selecionada, após escolher os melhores parâmetros de cada um dos modelos que serão testados, são gerados 1000 amostras diferentes de cada uma das técnicas de balanceamento e para cada modelo e gerados os resultados das métricas nos mil casos. O melhor Modelo Supervisionado é aquele que possui a maior média G-Mean dentre todos. Além disso, será analisado a estabilidade dos modelos treinados verifi-

cando a variância do G-Mean de cada método, assegurando assim a estabilidade do modelos, tendo em vista que a menor variância implica em uma dispersão em relação a média dos dados.

No Gráfico 4.6 há o resultado da Média e da Variância do G-Mean de cada um dos Modelos testados no estudo para cada Técnica de Balanceamento e as tabelas 4.4 e 4.5 mostram as estimativas representadas no Gráfico 4.6. O que se espera de um bom desempenho da previsão para os Modelos de Aprendizado de Máquina testados, é uma alta Média de G-Mean, para mostrar um bom desempenho da métrica do modelo, e uma baixa Variância, para mostrar a estabilidade do modelo. De modo geral, todos os modelos apresentaram uma baixa variância e isso quer dizer que os resultados *out-of-sample* são estáveis mesmo quando os modelos são estimados várias vezes. A estabilidade do modelo garante consistência para o resultado, isso quer dizer que o modelo consegue se replicável para várias amostras sem perder o poder de classificação. Variâncias altas significam valores de G-Mean diferindo muito ao longo as amostra, o que geraria um modelo com pouca consistência e com pouca replicabilidade.

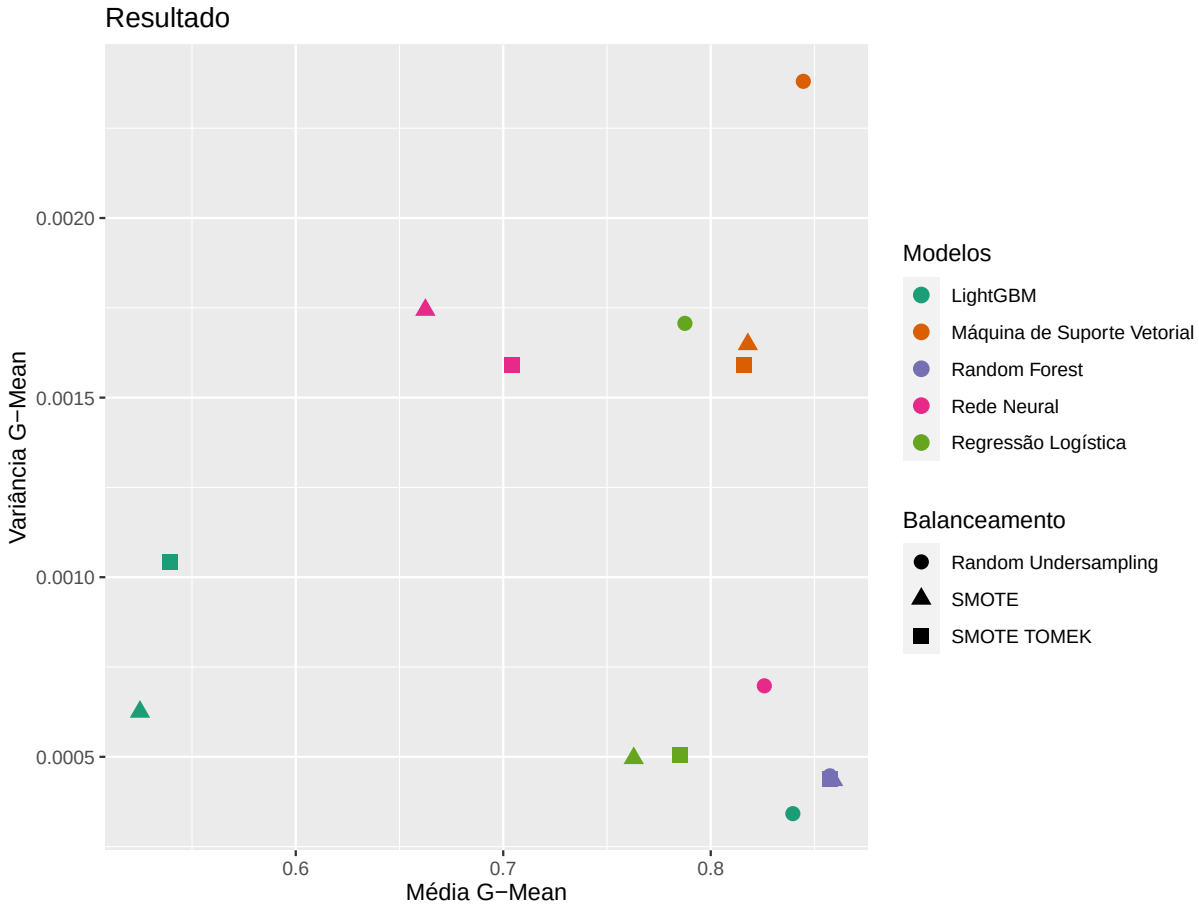

Figura 4.6 – Resultado *Geometric Mean Score*

Modelos	***Random Undersampling***	**SMOTE**	**SMOTETomek**
Regressão Logística	0,788	0,763	0,785
Random Forest	0,857	0,859	0,858
Máquina de Suporte Vetorial	0,845	0,818	0,816
LightGBM	0,840	0,525	0,539
Rede Neural	0,825	0,662	0,701

Tabela 4.4 – Média do *Geometric Mean Score*

Modelos	***Random Undersampling***	**SMOTE**	**SMOTETomek**
Regressão Logística	0,0017	0,0005	0,0005
Random Forest	0,0005	0,0004	0,0004
Máquina de Suporte Vetorial	0,0024	0,0016	0,0016
LightGBM	0,0003	0,0006	0,0010
Rede Neural	0,0006	0,0017	0,0016

Tabela 4.5 – Variância do *Geometric Mean Score*

O modelo de Regressão Logística obteve média de G-Mean de 0,788, 0,763 e 0,785 e Variâncias de 0,0017, 0,0005 e 0,0005, para o *Random Undersampling*, SMOTE e SMOTETomek, respectivamente. Por mais que *Random UnderSampling* tenha apresentado a maior média de G-Mean, a melhor técnica para este modelo, segundo a perspectiva da estabilidade, é o SMOTETomek, por apresentar uma menor variância.

O modelo de *Random Forest* obteve média de G-Mean de 0,857, 0,859 e 0,858 e Variância de 0,0005, 0,0004 e 0,0004 por não apresentar tanta diferença na variância o *Random Forest* é o modelo que apresentou maior estabilidade. A Técnica de Balanceamento que apresentou melhor resultado para o *Random Forest* foi o SMOTE, por apresentar maior G-Mean e a menor variância entre as outras duas técnicas.

O modelo de Máquina de Suporte Vetorial obteve média de G-Mean de 0,845, 0,818 e 0,816 para *Random Undersampling*, SMOTE e SMOTETomek, com variâncias de 0,0024, 0,0016 e 0,0016. A Técnica de Balanceamento que apresentou o melhor resultado, segundo a lógica de estabilidade que está sendo considerada neste estudo é o SMOTE, que apresentou o melhor G-Mean e baixa variância.

O modelo LightGBM obteve média de G-Mean de 0,840, 0,524 e 0,539 para *Random Undersampling*, SMOTE e SMOTETomek, com variâncias de 0,0003, 0,0006 e 0,0010. A Técnica de Balanceamento que apresentou melhor resultado é do *Random Undersampling* que apresentou melhor G-Mean e menos instabilidade devido a menor variância.

O modelo de Rede Neural obteve média de G-Mean de 0,825, 0,662 e 0,701 para *Random Undersampling*, SMOTE e SMOTETomek, com variâncias de 0,0006, 0,0017 e 0,0016. A Técnica de Balanceamento que apresentou melhor resultado é do *Random Undersampling* que apresentou melhor G-Mean e menos instabilidade devido a menor variância.

Os resultados dos Modelos mostram que *Random Forest* apresentou menor variabilidade no G-Mean entre os modelos para as três técnicas de Balanceamento, se mostrando como modelo mais estável. Por ser um modelo de Árvore, o *Random Forest* consegue discriminar as variáveis mais importantes e gerar resultados mais consistentes para dados desbalanceados. Além disso, o *Random Forest* apresentou as maiores médias de G-Mean, então além de mais estável o modelo ainda apresenta o melhor desempenho de previsão. Observando o contexto que foi estudado e as Técnicas de Balanceamento que foram testadas, recomenda-se o uso do *Random Forest* como para detectar e prever os indício de Fraude em Fundos de Pensão.

Observando as técnicas de Balanceamento que foram aplicadas, o SMOTE foi a técnica que obteve variâncias menores, isto é maior estabilidade, em mais de um caso do estudo. O SMOTE e o SMOTETomek são técnicas de *Oversampling* que imputam dados novos na base baseado no conjunto total. Isso faz com que os dados sejam pouco diversos do conjunto original, porque a técnica gera as novas observações de acordo com o comportamento dos dados originais. O fato das informações geradas serem semelhantes à base de de dados faz com que os resultados tenham a variância baixa para Modelos com uma quantidade menor de parâmetros.Já para o LightGBM e a Rede Neural, por serem modelos mais complexos e com maior quantidade de parâmetros, a falta de novos dados diversos fazem com que estes não apresentem bons resultados para o SMOTE e o SMOTETomek. Logo, para esses modelos, recomenda-se o uso de técnicas de balanceamento de ajustam as proporções sem imputar dados novos, como *Random Undersampling* que faz o balanceamento retirando dados das classes majoritárias.

4.4 Combinação de Modelos

Buscando novas possíveis interpretações, foi realizada uma análise de blendagem (*ensemble*) dos modelos para verificar possibilidades de combinação entre os modelos para melhorar os resultados. Esta análise foi denominada de Teste de Stress e foi interpretada da seguinte forma: com os melhores parâmetros de cada um dos modelos treinados, foi realizada mais uma rodada de previsão entretanto, dessa vez em toda a base de dados sem separação entre treinamento (*in-sample*) e validação (*out-of-sample*). O objetivo em realizar uma previsão que fosse global é analisar se a previsão gerada pelos modelos é semelhante à realidade informada pelos dados, ou seja, com

os modelos já treinados, agora é necessário observar como estes se comportam para a base completa.

Cada resultado gerado por cada modelo é comparado com os resultados reais e, assim, é calculado o G-Mean da previsão comparado à realidade. A evidência do Indício de Fraude é marcada como “Auto de Infração” na base de dados original, ou seja, caso seja encontrado algum Indício de Fraude para um Fundo de Pensão em um mês específico é rotulado na base de dados como “Auto de Infração”, conforme explicado pelos Auditores da PREVIC. Cada modelo estima uma previsão do Fundo de Pensão dentro do Período de Tempo, onde cada linha representa o Fundo no mês.

Após isso, o resultado real é comparado na situação em que dois modelos previram, isto é, é comparado o resultado real quando, por exemplo, a Regressão Logística e o *Random Forest* obtiveram a mesma previsão. Foram combinados os modelos dois a dois, três a três e quatro a quatro e casos em todos os modelos e, assim, foram geradas 26 possíveis combinações entre os 5 modelos testados. Os Gráficos 4.7, 4.8, 4.9 a seguir mostram os resultados para cada umas das Técnicas de Balanceamento baseado no G-Mean de cada combinação comparado ao resultado real. O significado de cada sigla nas legendas dos Gráficos 4.7, 4.8, 4.9 estão na Tabela 6.4 no Capítulo 6 (Apêndice).

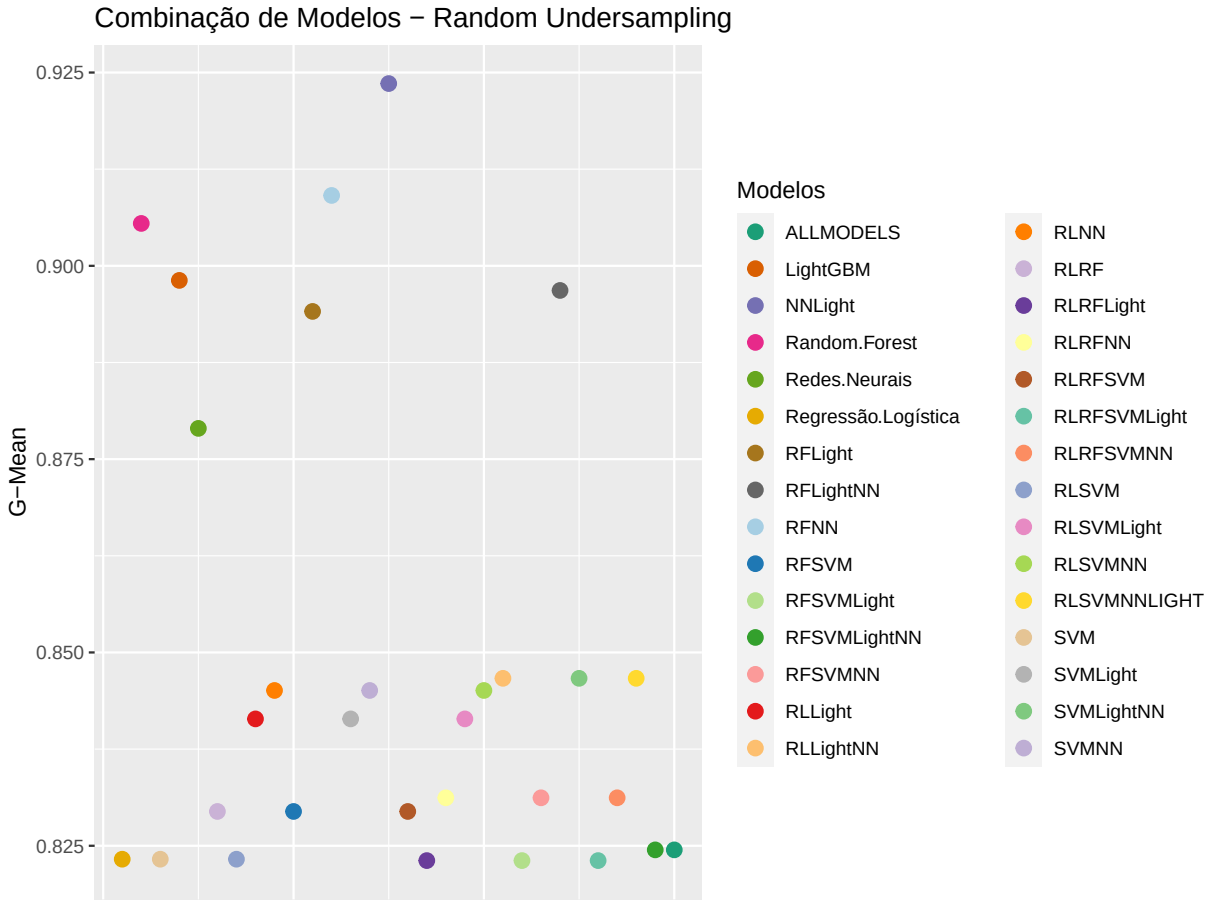

Figura 4.7 – Resultado Teste de Stress - *Random Undersampling*

Para o *Random Undersampling*, a combinação que apresentou maior G-Mean foi a Rede Neural e LightGBM com valor de 0.923, seguido pela combinação de Random Forest e Rede Neural com 0.910 e Random Forest com 0.904. O resultado da junção das previsões da Rede Neural e do LightGBM se aproxima mais do resultado real dos Indícios de Fraude. Nos resultados das previsões, que são apresentados na Seção 4.3, o *Random Undersampling* foi a técnica que apresentou menor variância para a Rede Neural e LightGBM se mostrando mais estável para modelos com mais parâmetros. É possível concluir que por mais que a previsão do LightGBM não tenha apresentado resultados tão bons na análise *out-of-sample* a combinação com a Rede Neural compensa o resultado, melhorando a previsão para a base completa.

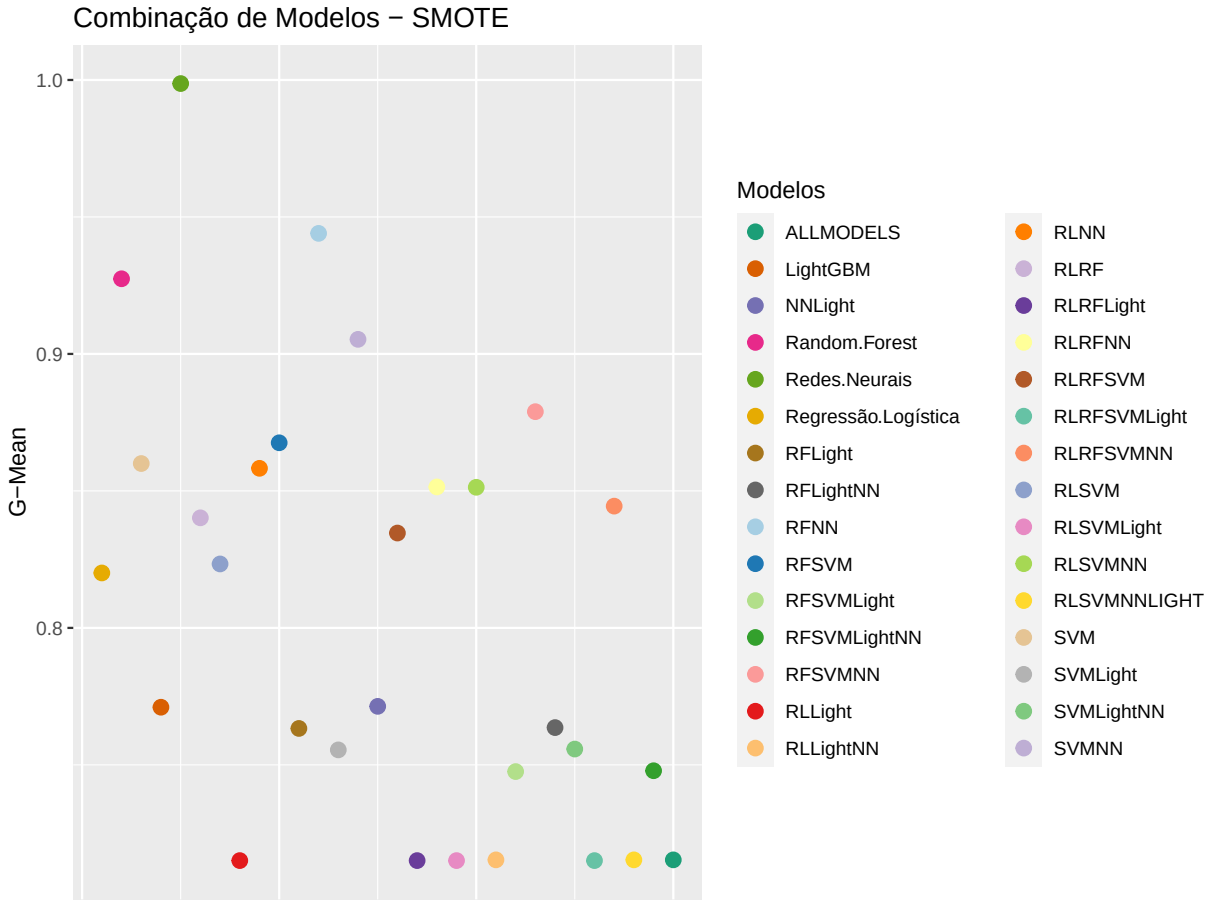

Figura 4.8 – Resultado Teste de Stress - SMOTE

Para o SMOTE, o modelo que apresenta maior G-Mean é o modelo Redes Neurais com 0.998, seguido da combinação do *Random Forest* com Rede Neural com 0.944 e *Random Forest* com 0.927. O resultado da comparação da Rede Neural com os dados reais mostram que o modelo se ajustou aos dados causando um *overfitting*, quando isso ocorre, há duas interpretações possíveis que podem ser feitas: a primeira é que o modelo consegue fazer a previsão assertiva dos dados disponíveis, porém não consegue generalizar para caso novos casos sejam acrescentados à base. Contudo, com a combinação do *Random Forest* e da Rede Neural, o problema do *overfitting* é sanado e mesmo assim o G-Mean se mantém com valor alto.

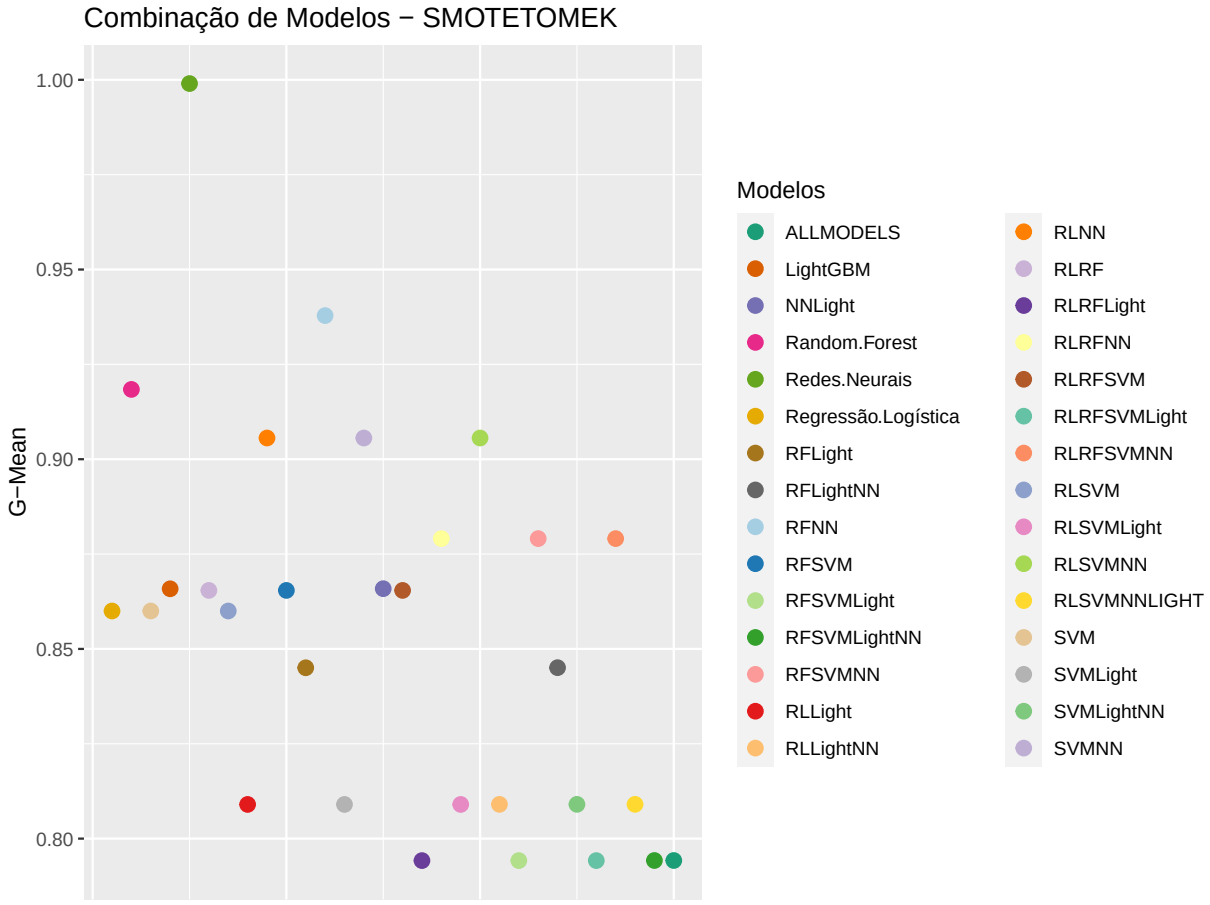

Figura 4.9 – Resultado Teste de Stress - SMOTETomek

Para o SMOTETomek, o modelo que apresenta maior G-Mean é o modelo Redes Neurais com 0.999, seguido da combinação do *Random Forest* com Rede Neural com 0.937 e *Random Forest* com 0.918, assim como a Técnica SMOTE. A interpretação para o SMOTETomek é semelhante a interpretação do SMOTE principalmente porque uma técnica é derivada da outra, trazendo resultados semelhantes para todas as combinações de modelos.

É importante frisar algumas interpretações que podem ser feitas a partir da análise com a combinação dos modelos. A junção de modelos que apresentaram resultados ruins com modelos que apresentaram resultados bons fez com que o desempenho dos modelos ruis melhorassem consideravelmente. Isso fica claro nos Gráficos 4.7, 4.8, 4.9 quando observamos os modelos combinados com Rede Neural e *Random Forest*. Por mais que a Rede Neural tenha uma tendência ao *overfitting*, quando combinado com modelos com menor parâmetros, o modelo se torna menos ajustado e, com isso, garante resultados mais adaptáveis a outras realidades.

Em relação as Técnicas de Balanceamento Utilizadas, o *Random Undersampling*,

assim como nos Resultados apresentados no Sessão 4.3, mostrou-se mais interessante de ser utilizado para modelos de Rede Neural e *LightGBM*. O SVM e a Regressão Logística não obtiveram bons resultados, entretanto quando combinados com Rede Neural, *Random Forest* ou *LightGBM* apresentaram melhora no desempenho. Para o SMOTE e o SMOTETomek, a melhora do desempenho dos modelos que não obtiveram bons resultados também é possível de observar com estas duas técnicas. Este tipo de análise é relevante para a área de estudo de Modelagem com Dados Desbalanceados porque traz novas possibilidades de aplicação dos Modelos e novas perspectivas perante modelos super-ajustados, como é o caso da Rede Neural, com SMOTE e SMOTETomek, que se torna mais flexível à classificação com a junção com *Random Forest*, contudo sem perder desempenho.

5 CONCLUSÃO

O estudo proposto teve como objetivo aplicar Modelos de Aprendizado de Máquina Supervisionados e de Técnicas de Balanceamento para prever fraudes em Fundos de Pensão. A base de dados utilizada foi cedida pela Superintendência Nacional de Previdência Complementar (PREVIC) com fim de auxiliar em novas perspectivas e possibilidades de metodologias para detecção de indícios de fraude, bem como modernização do processo realizado pelos auditores.

A detecção de fraudes em Fundos de Pensão possui um papel importante não somente para evitar perdas financeiras dos cotistas dos fundos, mas também auxiliar na fiscalização de Fundos de Previdência Complementar e, dessa forma, impedir atos de corrupção nessa área. Detectar Fraudes envolve lidar com eventos raros e, por consequência, com bases de dados altamente desbalanceadas. Um dos principais problemas de detecção de eventos raros é lidar com generalização dos modelos que, ao utilizar Aprendizado Indutivo, generaliza as classes majoritárias. Por isso, modelos tradicionais de classificação não são recomendados para dados desbalanceados, uma vez que os resultados da classificação podem não condizer com a realidade dos dados. O modelo tende a generalizar as classes majoritárias e, com isso, os resultados para as classes minoritárias apresentam resultados distorcidos. Isso acontece porque, em geral, o processo de aprendizado leva em consideração as classes principais dos dados para generalização e, assim, produz uma previsão que explique apenas a maioridade dos dados mas não necessariamente as classes minoritárias (HAIXIANG et al., 2017). Além dos problema de generalização dos modelos, a métrica mais utilizada na literatura da área, a acurácia, não avalia o desempenho dos modelos corretamente porque também tende para as classes majoritárias por ser uma métrica de avaliação global.

Com finalidade de solucionar as dificuldades envolvidas na aplicação de Modelos Supervisionados em dados desbalanceados, este estudo realizou uma série de pré-processamentos na base de dados. A dificuldade principal a ser sanada com o

pré-processamento foi devido à grande quantidade de *features* do conjunto de dados. Para retirar os possíveis ruídos que poderiam interferir na classificação e elencar as variáveis que seriam mais representativas para a base de dados, realizou-se em primeiro lugar uma regressão tipo LASSO e Ridge. Entretanto os parâmetros do modelo não apresentaram valores estatisticamente significantes e como os resultados das regressões não foram significativos, decidiu-se usar Análise de Componentes Principais (PCA) para diminuir a dimensão original da base de dados e concentrar a representatividade de várias variáveis em poucos componentes sem necessariamente perder muito da informação disponível. O resultado gerou 16 componentes principais que se concentraram nas variáveis da Dimensão de Investimentos. Para não concentrar o estudo somente em uma das dimensões, decidiu-se unir os resultados do PCA e das *dummies* da Dimensão de Dirigentes bem como da Dimensão de características do Fundo de Pensão. Nesta etapa, entendeu-se a necessidade do pré-processamento em dados desbalanceados devido a importância de um conjunto de dados que seja representativo e, além disso, não exija tanto tempo de processamento, devido a complexidade computacional dos modelos.

Após o pré-processamento, foi realizado um teste com as Métricas de Avaliação para selecionar qual métrica seria melhor para avaliar o desempenho em dados desbalanceados tendo em vista que o uso de métricas de avaliação inapropriadas é um dificuldade principal elencada por (WEISS, 2004). A métrica selecionada foi o*Geometric Mean Score* o qual é uma média geométrica entre a Sensibilidade (*Recall*) e a Especificidade. Selecionar a métrica é uma importante etapa do processo por se tratar da forma de avaliação do modelo. Com o teste efetuado, entende-se que a Acurácia não é uma métrica adequada para ser utilizada em dados desbalanceados por enviesar os resultados. Para observar o desempenho real, recomenda-se o uso de medidas que se relacionem com os Verdadeiros Positivos e Verdadeiros Negativos da Matriz de Confusão e isso foi abarcado com a métrica *Geometric Mean Score*.

Com a seleção da Métrica de Avaliação, foi realizado um estudo para escolha da melhor técnica de balanceamento para utilizar nas bases. O teste piloto foi feito com uma Regressão Logística e na base utilizada não foi feito nenhum pré-processamento com objetivo de testar como as técnicas de balanceamento se comportariam com os dados brutos. Foram testadas várias proporções de desbalanceamento entre as base de treinamento e base de teste para verificar a estabilidade dos métodos. As técnicas que obtiveram melhor resultado foram *Random Undersampling*, SMOTE e SMOTETomek e, por isso, foram as selecionadas para ser aplicadas nas base final.

O *Random Undersampling* é uma técnica de amostragem para diminuir a incidência da classe majoritária da base de dados e entre as técnicas de balanceamento testadas é a que possui a intuição mais simples de aplicação. Mesmo sendo uma técnica simples e, por vezes, a primeira opção de análise, o *Random Undersampling* se mostrou a melhor ferramenta entre as técnicas de *undersampling*, sendo a única entre as três selecionadas neste estudo seguindo esta lógica de amostragem, tendo um bom desempenho com os modelos com mais esparsos como LightGBM e Redes Neurais. A diminuição da quantidade de dados das classes majoritárias permite que o modelo se ajuste com mais consistência às classes que não foram re-amostradas, isto é, às classes que eram minoritárias antes da aplicação da técnica. Uma dificuldade encontrada é a tendência de "super-ajuste" dos modelos, ou seja, *overfitting*. Uma possível solução encontrada é a blendagem com outros modelos, como é visto com a união do *Random Forest* com a Rede Neural na Figura 4.7 na Sessão 4.4.

O SMOTE e SMOTETomek são técnicas de *oversampling* que aumentam a quantidade de dados das classes minoritárias na amostra. As duas técnicas são similares porque uma é derivada da outra, a diferença é que o SMOTETomek possui uma aprimoramento de cálculo com o uso da Tomek Links. São técnicas mais complexas e bastantes difundidas da literatura (RAHIM et al., 2019). O desempenho destas duas técnicas, tanto no sentido de previsão como estabilidade, foi melhor para Regressão Logística, *Random Forest* e Máquina de Suporte Vetorial. Como o SMOTE e o SMOTETomek inserem novos dados a partir das características das classes minoritárias, isso gera pouca similaridade entre bases fazendo com que estas técnicas funcionem melhor em modelos que fazem o processo de discriminação das classes utilizando amostras diferentes, como é o caso do *Random Forest* e em modelos que analisam a distância de característica dos dados como é exemplo da Regressão Logística e Máquina de Suporte Vetorial.

Na previsão, foram aplicados 5 modelos de Aprendizado de Máquina Supervisionados: Regressão Logística, *Random Forest*, Máquina de Suporte Vetorial, LightGBM e Redes Neurais. Como já foi citado anteriormente, foram considerados dois fatores para avaliar o desempenho do modelos, o G-Mean e a estabilidade, representada pela variância. Seguindo esta lógica, o melhor modelo dentre os modelos estudados foi o *Random Forest* utilizando SMOTE com 0,859 de G-Mean e 0,0004 de variância. O *Random Forest* apresentou as menores variâncias para as três técnicas de balanceamento de dados se mostrando também o modelo mais estável entre os 5 modelos. O modelo de Máquina de Suporte Vetorial apresentou as maiores variâncias se mostrando

o modelo mais instável, isto é, com menor consistência.

Em relação às Técnicas de Balanceamento, o *Random Undersampling* se mostrou mais instável por apresentar variâncias mais altas, por mais que apresente valores mais altos de G-Mean. O SMOTE e o SMOTETomek se mostraram mais estáveis para variância e o SMOTETomek apresentou melhores valores para o G-Mean.

Por fim, foi realizado uma análise final com a combinação dos modelos testados com a base completa e este resultados foram comparados com o resultado real base de indício de fraude. Com esta análise pode-se inferir que os modelos que apresentam mais parâmetros tem uma tendência a um super ajuste de generalização (*overfitting*) para os dados utilizados e uma solução proposta é o uso de combinação de outros modelos parar flexibilizar as análises, aumentando a generalização sem perder o desempenho.

Como abordado anteriormente, o objetivo deste estudo foi utilizar Modelos de Aprendizado de Máquina Supervisionados e Técnicas de Balanceamento para prever indício de fraudes em fundos de pensão. A questão principal envolvida foi solucionar o problema de bases desbalanceadas para realizar a previsão e auxiliar o processo de decisão da investigação do comportamento fraudulento ou não dos Fundos de Pensão analisados pela PREVIC.

Dado isso, algumas conclusões podem ser tomadas com respeito aos testes e análises realizados. Em primeiro lugar, a base de dados utilizada no processo de modelagem necessita de variáveis que sejam representativas e com pouco ruído para facilitar a discriminação das classes, além disso, como parte do pré-processamento da base de dados, é preciso aplicar técnicas para ajustar os modelos a lidar com diferente proporção entre as classes, trazendo uma solução para o balanceamento. A base utilizada neste estudo possuía uma proporção de 99,5% de não ocorrência para 0,5% de ocorrência de evento, neste caso o ajuste nas proporções foi de extrema importância para os resultados. Em segundo lugar, é necessário um olhar minucioso para métrica de avaliação do modelo com objetivo de mensurar o desempenho da melhor forma possível e de acordo com o objetivo da análise e gerenciais. Em terceiro lugar, é preciso observar não somente os valores das métricas, mas também como estas se comportam quando são testadas com novas amostras para verificar a estabilidade do resultado. E, por fim, ainda há a possibilidade de melhoria dos modelos utilizando-os de forma combinada.

Para fins de auxílio na Tomada de Decisão da PREVIC, este estudo recomenda

a utilização do *Random Forest* como Modelo de Aprendizado de Máquina, ajustando o desbalanceamento da base com o SMOTE. Acredita-se que com uso deste modelo, o processo de investigação pode ser modernizado e em certo ponto automatizado, além disso, o modelo pode proporcionar novos *insights* de investigação melhorando a abordagem já utilizada pelos auditores.

O presente estudo possui limitações de análise que podem ser intuições para trabalhos futuros dentro da área de pesquisa. A primeira limitação é a falta de aplicação em Modelos Não-Supervisionados tratando da abordagem Proativa em relação aos dados, bem como de outros Modelos Supervisionados mais recentes, que trariam outra ótica para os resultados e para o comportamento das variáveis. Outra limitação, seria a falta de teste para outras Técnicas de Balanceamento, no estudo apenas três foram utilizadas com fim de comparação, entretanto seria interessante uma análise com outras técnicas para expansão das possibilidades. Finalmente, não foram testados modelos específicos do campo de detecção de *outlier*, dado que a proporção entre as classes é extremamente desbalanceada, acredita-se que o uso de modelos característicos para *outlier* poderiam trazer novas abordagens e perspectivas úteis para a previsão e reconhecimentos de eventos raros.

6 APÊNDICE

Variável	Interpretação da Variável
PERIODO_PONDERADO	Período
TOTAL	Total da Carteira
retACOES1	Retorno das Ações no último ano
retDEPOSITO1	Retorno dos Depósitos no último ano
retDIREITO_CREDITORIO1	Retorno de Direito Creditório no último ano
retEMPRESTIMO1	Retorno dos Empréstimos no último ano
retFINANCIAMENTO_IMOBILIARIO1	Retorno em Financiamento Imobiliário no último ano
retIMOVEL1	Retorno de Investimento em Imóvel no último ano
retCOTASDEFUNDO1	Retorno de Investimentos em Cotas de Outros Fundos no último ano
retVL_PAGAR_VLRECEBER1	Retorno dos Valores a Pagar e Receber no último ano
retOPERACAOCOMPROMISSADA1	Retorno em Operação Compromissada no último ano
retTITULOPRIVADO1	Retorno em Títulos Privados no último ano
retTITULOPUBLICO1	Retorno em Títulos Públicos no último ano
retACOES2	Retorno das Ações dos últimos 2 anos
retDEPOSITO2	Retorno dos Depósitos dos últimos 2 anos
retDIREITO_CREDITORIO2	Retorno de Direito Creditório dos últimos 2 anos
retEMPRESTIMO2	Retorno dos Empréstimos dos últimos 2 anos
retFINANCIAMENTO_IMOBILIARIO2	Retorno em Financiamento Imobiliário dos últimos 2 anos
retIMOVEL2	Retorno de Investimentos em Imóveis em dos últimos 2 anos

retCOTASDEFUNDO2	Retorno de Investimentos em Cotas de Outros Fundos nos últimos 2 anos
retVL_PAGAR_VLRECEBER2	Retorno dos Valores a Pagar e Receber nos últimos 2 anos
retOPERACAOCOMPROMISSADA2	Retorno em Operação Compromissada nos últimos 2 anos
retTITULOPRIVADO2	Retorno em Títulos Privados nos últimos 2 anos
retTITULOPUBLICO2	Retorno em Títulos Públicos nos últimos 2 anos
retACOES3	Retorno das Ações dos últimos 3 anos
retDEPOSITO3	Retorno dos Depósitos dos últimos 3 anos
retDIREITO_CREDITORIO3	Retorno de Direito Creditório dos últimos 3 anos
retEMPRESTIMO3	Retorno dos Empréstimos dos últimos 3 anos
retFINANCIAMENTO_IMOBILIARIO3	Retorno em Financiamento Imobiliário dos últimos 3 anos
retIMOVEL3	Retorno de Investimentos em Imóveis em dos últimos 3 anos
retCOTASDEFUNDO3	Retorno de Investimentos em Cotas de Outros Fundos nos últimos 3 anos
retVL_PAGAR_VLRECEBER3	Retorno dos Valores a Pagar e Receber nos últimos 3 anos
retOPERACAOCOMPROMISSADA3	Retorno em Operação Compromissada nos últimos 3 anos
retTITULOPRIVADO3	Retorno em Títulos Privados nos últimos 3 anos
retTITULOPUBLICO3	Retorno em Títulos Públicos nos últimos 3 anos
retACOES4	Retorno das Ações dos últimos 4 anos
retDEPOSITO4	Retorno dos Depósitos dos últimos 4 anos
retDIREITO_CREDITORIO4	Retorno de Direito Creditório dos últimos 4 anos
retEMPRESTIMO4	Retorno dos Empréstimos dos últimos 4 anos
retFINANCIAMENTO_IMOBILIARIO4	Retorno em Financiamento Imobiliário dos últimos 4 anos

retIMOVEL4	Retorno de Investimentos em Imóveis dos últimos 4 anos
retCOTASDEFUNDO4	Retorno de Investimentos em Cotas de Outros Fundos nos últimos 4 anos
retVL_PAGAR_VLRECEBER4	Retorno dos Valores a Pagar e Receber nos últimos 4 anos
retOPERACAOCOMPROMISSADA4	Retorno em Operação Compromissada nos últimos 4 anos
retTITULOPRIVADO4	Retorno em Títulos Privados nos últimos 4 anos
retTITULOPUBLICO4	Retorno em Títulos Públicos nos últimos 4 anos
retACOES5	Retorno das Ações dos últimos 5 anos
retDEPOSITO5	Retorno dos Depósitos dos últimos 5 anos
retDIREITO_CREDITORIO5	Retorno de Direito Creditório dos últimos 5 anos
retEMPRESTIMO5	Retorno dos Empréstimos dos últimos 5 anos
retFINANCIAMENTO_IMOBILIARIO5	Retorno em Financiamento Imobiliário dos últimos 5 anos
retIMOVEL5	Retorno de Investimentos em Imóveis em dos últimos 5 anos
retCOTASDEFUNDO5	Retorno de Investimentos em Cotas de Outros Fundos nos últimos 5 anos
retVL_PAGAR_VLRECEBER5	Retorno dos Valores a Pagar e Receber nos últimos 5 anos
retOPERACAOCOMPROMISSADA5	Retorno em Operação Compromissada nos últimos 5 anos
retTITULOPRIVADO5	Retorno em Títulos Privados nos últimos 5 anos
retTITULOPUBLICO5	Retorno em Títulos Públicos nos últimos 5 anos
lagPT1	Presença de Dirigente Filiado ao PT no último ano
lagAVANTE1	Presença de Dirigente Filiado ao AVANTE no último ano

lagCIDADANIA1	Presença de Dirigente Filiado ao CIDADANIA no último ano
lagMDB1	Presença de Dirigente Filiado ao MDB no último ano
lagPCdoB1	Presença de Dirigente Filiado ao PCdoB no último ano
lagPMN1	Presença de Dirigente Filiado ao PMN no último ano
lagPMB1	Presença de Dirigente Filiado ao PMB no último ano
lagPSDB1	Presença de Dirigente Filiado ao PSDB no último ano
lagPDT1	Presença de Dirigente Filiado ao PDT no último ano
lagPL1	Presença de Dirigente Filiado ao PL no último ano
lagNOVO1	Presença de Dirigente Filiado ao NOVO no último ano
lagPRTN1	Presença de Dirigente Filiado ao PRTN no último ano
lagPROS1	Presença de Dirigente Filiado ao PROS no último ano
lagPSC1	Presença de Dirigente Filiado ao PSC no último ano
lagPSD1	Presença de Dirigente Filiado ao PSD no último ano
lagPSL1	Presença de Dirigente Filiado ao PSL no último ano
lagPSEL1	Presença de Dirigente Filiado ao PSEL no último ano
lagPSTU1	Presença de Dirigente Filiado ao PSTU no último ano
lagPSB1	Presença de Dirigente Filiado ao PSB no último ano

lagPTB1	Presença de Dirigente Filiado ao PTB no último ano
lagPV1	Presença de Dirigente Filiado ao PV no último ano
IDADEFUNDO	Idade do Fundo (em meses)
Instituidor	Patrocinador Instituidor
Privada	Patrocinador Privado
Pública Estadual	Patrocinador Público Estadual
Pública Federal	Patrocinador Público Federal
Pública Municipal	Patrocinador Público Municipal

Tabela 6.1 – Dicionário de Variáveis

Variável	**Beta**
Intercepto	0,0007
Período	0
Total da Carteira	0
Retorno das Ações no último ano	0
Retorno dos Depósitos no último ano	0
Retorno de Direito Creditório no último ano	0
Retorno dos Empréstimos no último ano	0
Retorno em Financiamento Imobiliário no último ano	0
Retorno de Investimento em Imóvel no último ano	0
Retorno de Investimentos em Cotas de Outros Fundos no último ano	0
Retorno dos Valores a Pagar e Receber no último ano	0
Retorno em Operação Compromissada no último ano	0
Retorno em Títulos Privados no último ano	0
Retorno em Títulos Públicos no último ano	0
Retorno das Ações dos últimos 2 anos	0
Retorno dos Depósitos dos últimos 2 anos	0
Retorno de Direito Creditório dos últimos 2 anos	0
Retorno dos Empréstimos dos últimos 2 anos	0
Retorno em Financiamento Imobiliário dos últimos 2 anos	0
Retorno de Investimentos em Imóveis em dos últimos 2 anos	0

Retorno de Investimentos em Cotas de Outros Fundos nos últimos 2 anos	0
Retorno dos Valores a Pagar e Receber nos últimos 2 anos	0
Retorno em Operação Compromissada nos últimos 2 anos	0
Retorno em Títulos Privados nos últimos 2 anos	0
Retorno em Títulos Públicos nos últimos 2 anos	0
Retorno das Ações dos últimos 3 anos	0
Retorno dos Depósitos dos últimos 3 anos	0
Retorno de Direito Creditório dos últimos 3 anos	0
Retorno dos Empréstimos dos últimos 3 anos	0
Retorno em Financiamento Imobiliário dos últimos 3 anos	0
Retorno de Investimentos em Imóveis em dos últimos 3 anos	0
Retorno de Investimentos em Cotas de Outros Fundos nos últimos 3 anos	0
Retorno dos Valores a Pagar e Receber nos últimos 3 anos	0
Retorno em Operação Compromissada nos últimos 3 anos	0
Retorno em Títulos Privados nos últimos 3 anos	0
Retorno em Títulos Públicos nos últimos 3 anos	0
Retorno das Ações dos últimos 4 anos	0
Retorno dos Depósitos dos últimos 4 anos	0
Retorno de Direito Creditório dos últimos 4 anos	0
Retorno dos Empréstimos dos últimos 4 anos	0
Retorno em Financiamento Imobiliário dos últimos 4 anos	0
Retorno de Investimentos em Imóveis em dos últimos 4 anos	0
Retorno de Investimentos em Cotas de Outros Fundos nos últimos 4 anos	0
Retorno dos Valores a Pagar e Receber nos últimos 4 anos	0
Retorno em Operação Compromissada nos últimos 4 anos	0
Retorno em Títulos Privados nos últimos 4 anos	0
Retorno em Títulos Públicos nos últimos 4 anos	0
Retorno das Ações dos últimos 5 anos	0
Retorno dos Depósitos dos últimos 5 anos	0
Retorno de Direito Creditório dos últimos 5 anos	0
Retorno dos Empréstimos dos últimos 5 anos	0
Retorno em Financiamento Imobiliário dos últimos 5 anos	0

Retorno de Investimentos em Imóveis em dos últimos 5 anos	0
Retorno de Investimentos em Cotas de Outros Fundos nos últimos 5 anos	0
Retorno dos Valores a Pagar e Receber nos últimos 5 anos	0
Retorno em Operação Compromissada nos últimos 5 anos	0
Retorno em Títulos Privados nos últimos 5 anos	0
Retorno em Títulos Públicos nos últimos 5 anos	0
Presença de Dirigente Filiado ao PT no último ano	0
Presença de Dirigente Filiado ao AVANTE no último ano	0
Presença de Dirigente Filiado ao CIDADANIA no último ano	0
Presença de Dirigente Filiado ao MDB no último ano	0
Presença de Dirigente Filiado ao PCdoB no último ano	0
Presença de Dirigente Filiado ao PMN no último ano	0
Presença de Dirigente Filiado ao PMB no último	0
Presença de Dirigente Filiado ao PSDB no último ano	0
Presença de Dirigente Filiado ao PDT no último ano	0
Presença de Dirigente Filiado ao PL no último ano	0
Presença de Dirigente Filiado ao NOVO no último ano	0
Presença de Dirigente Filiado ao PRTN no último ano	0
Presença de Dirigente Filiado ao PROS no último ano	0
Presença de Dirigente Filiado ao PSC no último ano	0
Presença de Dirigente Filiado ao PSD no último ano	0
Presença de Dirigente Filiado ao PSL no último ano	0
Presença de Dirigente Filiado ao PSEL no último ano	0
Presença de Dirigente Filiado ao PSTU no último ano	0
Presença de Dirigente Filiado ao PSB no último ano	0
Presença de Dirigente Filiado ao PTB no último ano	0
Presença de Dirigente Filiado ao PV no último ano	0
Idade do Fundo (em meses)	0
Patrocinador Instituidor	0
Patrocinador Privado	0
Patrocinador Público Estadual	0
Patrocinador Público Federal	0
Patrocinador Público Municipal	0

Tabela 6.2 – Valores do Beta da Regressão LASSO

Variável	**Beta**
Intercepto	0,0048
Período	0
Total da Carteira	0
Retorno das Ações no último ano	0
Retorno dos Depósitos no último ano	0
Retorno de Direito Creditório no último ano	0
Retorno dos Empréstimos no último ano	0
Retorno em Financiamento Imobiliário no último ano	0
Retorno de Investimento em Imóvel no último ano	0
Retorno de Investimentos em Cotas de Outros Fundos no último ano	0
Retorno dos Valores a Pagar e Receber no último ano	0
Retorno em Operação Compromissada no último ano	0
Retorno em Títulos Privados no último ano	0
Retorno em Títulos Públicos no último ano	0
Retorno das Ações dos últimos 2 anos	0
Retorno dos Depósitos dos últimos 2 anos	0
Retorno de Direito Creditório dos últimos 2 anos	0
Retorno dos Empréstimos dos últimos 2 anos	0
Retorno em Financiamento Imobiliário dos últimos 2 anos	0
Retorno de Investimentos em Imóveis em dos últimos 2 anos	0
Retorno de Investimentos em Cotas de Outros Fundos nos últimos 2 anos	0
Retorno dos Valores a Pagar e Receber nos últimos 2 anos	0
Retorno em Operação Compromissada nos últimos 2 anos	0
Retorno em Títulos Privados nos últimos 2 anos	0
Retorno em Títulos Públicos nos últimos 2 anos	0
Retorno das Ações dos últimos 3 anos	0
Retorno dos Depósitos dos últimos 3 anos	0
Retorno de Direito Creditório dos últimos 3 anos	0
Retorno dos Empréstimos dos últimos 3 anos	0

Retorno em Financiamento Imobiliário dos últimos 3 anos	0
Retorno de Investimentos em Imóveis em dos últimos 3 anos	0
Retorno de Investimentos em Cotas de Outros Fundos nos últimos 3 anos	0
Retorno dos Valores a Pagar e Receber nos últimos 3 anos	0
Retorno em Operação Compromissada nos últimos 3 anos	0
Retorno em Títulos Privados nos últimos 3 anos	0
Retorno em Títulos Públicos nos últimos 3 anos	0
Retorno das Ações dos últimos 4 anos	0
Retorno dos Depósitos dos últimos 4 anos	0
Retorno de Direito Creditório dos últimos 4 anos	0
Retorno dos Empréstimos dos últimos 4 anos	0
Retorno em Financiamento Imobiliário dos últimos 4 anos	0
Retorno de Investimentos em Imóveis em dos últimos 4 anos	0
Retorno de Investimentos em Cotas de Outros Fundos nos últimos 4 anos	0
Retorno dos Valores a Pagar e Receber nos últimos 4 anos	0
Retorno em Operação Compromissada nos últimos 4 anos	0
Retorno em Títulos Privados nos últimos 4 anos	0
Retorno em Títulos Públicos nos últimos 4 anos	0
Retorno das Ações dos últimos 5 anos	0
Retorno dos Depósitos dos últimos 5 anos	0
Retorno de Direito Creditório dos últimos 5 anos	0
Retorno dos Empréstimos dos últimos 5 anos	0
Retorno em Financiamento Imobiliário dos últimos 5 anos	0
Retorno de Investimentos em Imóveis em dos últimos 5 anos	0
Retorno de Investimentos em Cotas de Outros Fundos nos últimos 5 anos	0
Retorno dos Valores a Pagar e Receber nos últimos 5 anos	0
Retorno em Operação Compromissada nos últimos 5 anos	0
Retorno em Títulos Privados nos últimos 5 anos	0
Retorno em Títulos Públicos nos últimos 5 anos	0
Presença de Dirigente Filiado ao PT no último ano	0
Presença de Dirigente Filiado ao AVANTE no último ano	0
Presença de Dirigente Filiado ao CIDADANIA no último ano	0

Presença de Dirigente Filiado ao MDB no último ano	0
Presença de Dirigente Filiado ao PCdoB no último ano	0
Presença de Dirigente Filiado ao PMN no último ano	0
Presença de Dirigente Filiado ao PMB no último ano	0
Presença de Dirigente Filiado ao PSDB no último ano	0
Presença de Dirigente Filiado ao PDT no último ano	0
Presença de Dirigente Filiado ao PL no último ano	0
Presença de Dirigente Filiado ao NOVO no último ano	0
Presença de Dirigente Filiado ao PRTN no último ano	0
Presença de Dirigente Filiado ao PROS no último ano	0
Presença de Dirigente Filiado ao PSC no último ano	0
Presença de Dirigente Filiado ao PSD no último ano	0
Presença de Dirigente Filiado ao PSL no último ano	0
Presença de Dirigente Filiado ao PSEL no último ano	0
Presença de Dirigente Filiado ao PSTU no último ano	0
Presença de Dirigente Filiado ao PSB no último ano	0
Presença de Dirigente Filiado ao PTB no último ano	0
Presença de Dirigente Filiado ao PV no último ano	0
Idade do Fundo (em meses)	0
Patrocinador Instituidor	0
Patrocinador Privado	0
Patrocinador Público Estadual	0
Patrocinador Público Federal	0
Patrocinador Público Municipal	0

Tabela 6.3 – Valores do Beta da Regressão Ridge

Combinação de Modelo	Sigla
Regressão Logística e Random Forest	RLRF
Regressão Logística e SVM	RLSVM
Regressão Logística e LightGBM	RLLight
Regressão Logística e Rede Neurais	RLNN
Random Forest e SVM	RFSVM
Random Forest e LightGBM	RFLight
Random Forest e Redes Neurais	RFNN
SVM e LightGBM	SVMLight
SVM e Redes Neurais	SVMNN
Redes Neurais e LightGBM	NNLight
Regressão Logística, Random Forest e SVM	RLRFSVM
Regressão Logística, Random Forest e LightGBM	RLRFLight
Regressão Logística, Random Forest e Redes Neurais	RLRFNN
Regressão Logística, SVM e LightGBM	RLSVMLight
Regressão Logística, SVM e Redes Neurais	RLSVMNN
Regressão Logística, LightGBM e Redes Neurais	RLLightNN
Random Forest, SVM e LightGBM	RFSVMLight
Random Forest, SVM e Redes Neurais	RFSVMNN
Random Forest, LightGBM e Redes Neurais	RFLightNN
SVM, LightGBM e Redes Neurais	SVMLightNN
Regressão Logística, Random Forest, SVM e LightGBM	RLRFSVMLight
Regressão Logística, Random Forest, SVM e Redes Neurais	RLRFSVMNN
Regressão Logística, SVM, Redes Neurais e LightGBM	RLSVMNNLIGHT
Random Forest, SVM, LightGBM e Redes Neurais	RFSVMLightNN
Todos os Modelos	ALLMODELS

Tabela 6.4 – Combinação dos Modelos x Siglas

REFERÊNCIAS BIBLIOGRÁFICAS

ABDELHAMID, D.; KHAOULA, S.; ATIKA, O. Automatic bank fraud detection using support vector machines. **The International Conference on Computing Technology and Information Management (ICCTIM)**. [S.l.], 2014. p. 10.

ABDI, H.; WILLIAMS, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010.

Agência Brasil. *Procuradora explica esquema de fraude em fundo de pensão da Petrobras*. 2018. Accessado em 28/02/2020. Disponível em: <http://agenciabrasil.ebc.com.br/geral/noticia/2018-11/procuradora-explica-esquema-de-fraude-em-fundo-de-pensao-da-petrobras>.

AHMED, M.; MAHMOOD, A. N.; ISLAM, M. R. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, Elsevier, v. 55, p. 278–288, 2016.

ALBIERO, B.; SANTOS, R.; UYRÁ, E.; VILARINO, R.; SILVA, J.; SOUZA, T.; VICENTE, R.; YAMOUNI, S. Employing gradient boosting and anomaly detection for prediction of frauds in energy consumption. **Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2019. p. 916–925.

ALBUQUERQUE, P. H. *Previsão de séries temporais financeiras por meio de máquinas de suporte vetorial e ondaletas*. [S.l.], 2014.

ALBUQUERQUE, P. H.; PENG, Y.; NAKANO, E.; SILVA, C. D.; BOSQUE, L. Probability of informed trading: a bayesian approach. *International Journal of Applied Decision Sciences*, 2019.

ARINO, K.; KIKUTA, Y. Classsim: Similarity between classes defined by misclassification ratios of trained classifiers. *arXiv preprint arXiv:1802.01267*, 2018.

BARBOZA, F.; KIMURA, H.; ALTMAN, E. Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, Elsevier, v. 83, p. 405–417, 2017.

BARMAN, S.; PAL, U.; SARFARAJ, M. A.; BISWAS, B.; MAHATA, A.; MANDAL, P. A complete literature review on financial fraud detection applying data mining techniques. *International Journal of Trust Management in Computing and Communications*, Inderscience Publishers (IEL), v. 3, n. 4, p. 336–359, 2016.

BARUA, S.; ISLAM, M. M.; YAO, X.; MURASE, K. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 26, n. 2, p. 405–425, 2012.

BATISTA, G. E.; BAZZAN, A. L.; MONARD, M. C. et al. Balancing training data for automated annotation of keywords: a case study. **WOB**. [S.l.: s.n.], 2003. p. 10–18.

BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, ACM New York, NY, USA, v. 6, n. 1, p. 20–29, 2004.

BBC News. *Insurance fraud 'costs UK £1.6bn'*. 2007. Accessado em 10/12/2019. Disponível em: <http://news.bbc.co.uk/2/hi/business/6636005.stm>.

BEYAN, C.; FISHER, R. Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition*, Elsevier, v. 48, n. 5, p. 1653–1672, 2015.

BHATTACHARYYA, S.; JHA, S.; THARAKUNNEL, K.; WESTLAND, J. C. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, Elsevier, v. 50, n. 3, p. 602–613, 2011.

Bitcoin. 2019. Accessed: 18/12/2019. Disponível em: <https://bitcoin.org/en/>.

BOLTON, R. J.; HAND, D. J. Statistical fraud detection: A review. *Statistical science*, JSTOR, p. 235–249, 2002.

BRANCO, P.; TORGO, L.; RIBEIRO, R. A survey of predictive modelling under imbalanced distributions. *arXiv preprint arXiv:1505.01658*, 2015.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

CAKIR, E.; VIRTANEN, T. Convolutional recurrent neural networks for rare sound event detection. *Deep Neural Networks for Sound Event Detection*, v. 12, 2019.

CASTRO, C. L.; BRAGA, A. de P. Artificial neural networks learning in roc space. **IJCCI**. [S.l.], 2009. p. 484–489.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, ACM, v. 41, n. 3, p. 15, 2009.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.

CHEN, C.; LIAW, A.; BREIMAN, L. et al. Using random forest to learn imbalanced data. *University of California, Berkeley*, v. 110, n. 1-12, p. 24, 2004.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 785–794.

CHI, H.-M.; ERSOY, O. K. Support vector machine decision trees with rare event detection. *International Journal of Smart Engineering System Design*, Taylor & Francis, v. 4, n. 4, p. 225–242, 2002.

CODY, C.; FORD, V.; SIRAJ, A. Decision tree learning for fraud detection in consumer energy consumption. **2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)**. [S.l.], 2015. p. 1175–1179.

Correio Braziliense. *Fraudes na Previdência Social somam R$ 5,5 bilhões em 16 anos*. 2019. Acessado em 06/12/2019. Disponível em: <https://www.correiobraziliense.com.br/app/noticia/brasil/2019/09/29/interna-brasil,792219/fraudes-na-previdencia-social-somam-r-5-5-bilhoes-em-16-anos.shtml>.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

CULKIN, R.; DAS, S. R. Machine learning in finance: the case of deep learning for option pricing. *Journal of Investment Management*, v. 15, n. 4, p. 92–100, 2017.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Random forests. In: *Ensemble machine learning*. [S.l.]: Springer, 2012. p. 157–175.

DOUZAS, G.; BACAO, F.; LAST, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. *Information Sciences*, Elsevier, v. 465, p. 1–20, 2018.

EFFENDY, V.; BAIZAL, Z. A. et al. Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. **2014 2nd International Conference on Information and Communication Technology (ICoICT)**. [S.l.], 2014. p. 325–330.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001.

FU, K.; CHENG, D.; TU, Y.; ZHANG, L. Credit card fraud detection using convolutional neural networks. **International Conference on Neural Information Processing**. [S.l.], 2016. p. 483–490.

GALAR, M.; FERNANDEZ, A.; BARRENECHEA, E.; BUSTINCE, H.; HERRERA, F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, IEEE, v. 42, n. 4, p. 463–484, 2011.

HAIXIANG, G.; YIJING, L.; SHANG, J.; MINGYUN, G.; YUANYUE, H.; BING, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, Elsevier, v. 73, p. 220–239, 2017.

HAN, H.; WANG, W.-Y.; MAO, B.-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. **International conference on intelligent computing**. [S.l.], 2005. p. 878–887.

HANIFAH, F. S.; WIJAYANTO, H.; KURNIA, A. Smotebagging algorithm for imbalanced dataset in logistic regression analysis (case: Credit of bank x). *Applied Mathematical Sciences*, v. 9, n. 138, p. 6857–6865, 2015.

HART, P. The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory*, Citeseer, v. 14, n. 3, p. 515–516, 1968.

HASANIN, T.; KHOSHGOFTAAR, T. The effects of random undersampling with simulated class imbalance for big data. **2018 IEEE International Conference on Information Reuse and Integration (IRI)**. [S.l.], 2018. p. 70–79.

HAYKIN, S. *Neural networks: a comprehensive foundation*. [S.l.]: Prentice Hall PTR, 1994.

HENRIQUE, B. M.; SOBREIRO, V. A.; KIMURA, H. Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, Elsevier, 2019.

HU, X.; CHEN, H.; ZHANG, R. Short paper: Credit card fraud detection using lightgbm with asymmetric error control. **2019 Second International Conference on Artificial Intelligence for Industries (AI4I)**. [S.l.], 2019. p. 91–94.

JANJUA, Z. H.; VECCHIO, M.; ANTONINI, M.; ANTONELLI, F. Irese: An intelligent rare-event detection system using unsupervised learning on the iot edge. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 84, p. 41–50, 2019.

KAISER, Ł.; NACHUM, O.; ROY, A.; BENGIO, S. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*, 2017.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in neural information processing systems**. [S.l.: s.n.], 2017. p. 3146–3154.

KING, G.; ZENG, L. Logistic regression in rare events data. *Political analysis*, Cambridge University Press, v. 9, n. 2, p. 137–163, 2001.

KIRKOS, E.; SPATHIS, C.; MANOLOPOULOS, Y. Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, Elsevier, v. 32, n. 4, p. 995–1003, 2007.

KIRLIDOG, M.; ASUK, C. A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences*, Elsevier, v. 62, p. 989–994, 2012.

KÖKNAR-TEZEL, S.; LATECKI, L. J. Improving svm classification on imbalanced data sets in distance spaces. **2009 Ninth IEEE International Conference on Data Mining**. [S.l.], 2009. p. 259–267.

KOU, Y.; LU, C.-T.; SIRWONGWATTANA, S.; HUANG, Y.-P. Survey of fraud detection techniques. **IEEE International Conference on Networking, Sensing and Control, 2004**. [S.l.], 2004. v. 2, p. 749–754.

KRAUSS, C.; DO, X. A.; HUCK, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, Elsevier, v. 259, n. 2, p. 689–702, 2017.

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, Springer, v. 5, n. 4, p. 221–232, 2016.

KUMAR, M.; THENMOZHI, M. Forecasting stock index returns using arima-svm, arima-ann, and arima-random forest hybrid models. *International Journal of Banking, Accounting and Finance*, Inderscience Publishers Ltd, v. 5, n. 3, p. 284–308, 2014.

LEE, J.-S. Auc4. 5: Auc-based c4. 5 decision tree algorithm for imbalanced data classification. *IEEE Access*, IEEE, v. 7, p. 106034–106042, 2019.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, JMLR. org, v. 18, n. 1, p. 559–563, 2017.

LI, C. Classifying imbalanced data using a bagging ensemble variation (bev). **Proceedings of the 45th annual southeast regional conference**. [S.l.: s.n.], 2007. p. 203–208.

LUQUE, A.; CARRASCO, A.; MARTÍN, A.; HERAS, A. de las. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, Elsevier, v. 91, p. 216–231, 2019.

MAALOUF, M.; HOMOUZ, D.; TRAFALIS, T. B. Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. *Computational Intelligence*, Wiley Online Library, v. 34, n. 1, p. 161–174, 2018.

MAALOUF, M.; SIDDIQI, M. Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, Elsevier, v. 59, p. 142–148, 2014.

MAALOUF, M.; TRAFALIS, T. B. Rare events and imbalanced datasets: an overview. *International Journal of Data Mining, Modelling and Management*, Inderscience Publishers, v. 3, n. 4, p. 375–388, 2011.

MAES, S.; TUYLS, K.; VANSCHOENWINKEL, B.; MANDERICK, B. Credit card fraud detection using bayesian and neural networks. **Proceedings of the 1st international naiso congress on neuro fuzzy technologies**. [S.l.: s.n.], 2002. p. 261–270.

MENDES, B. V. M. *Introdução à análise de eventos extremos*. [S.l.]: Editora E-papers, 2004.

MENSAH, C.; KLEIN, J.; BHULAI, S.; HOOGENDOORN, M.; MEI, R. van der. Detecting fraudulent bookings of online travel agencies with unsupervised machine learning. **International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems**. [S.l.], 2019. p. 334–346.

MISHRA, S. Handling imbalanced data: Smote vs. random undersampling. *Int. Res. J. Eng. Technol*, v. 4, n. 8, p. 317–320, 2017.

MOHAMMED, R.; RAWASHDEH, J.; ABDULLAH, M. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. **2020 11th International Conference on Information and Communication Systems (ICICS)**. [S.l.], 2020. p. 243–248.

MUCHLINSKI, D.; SIROKY, D.; HE, J.; KOCHER, M. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, Cambridge University Press, v. 24, n. 1, p. 87–103, 2016.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, Frontiers, v. 7, p. 21, 2013.

NGAI, E. W.; HU, Y.; WONG, Y. H.; CHEN, Y.; SUN, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, Elsevier, v. 50, n. 3, p. 559–569, 2011.

NGUYEN, H. M.; COOPER, E. W.; KAMEI, K. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, Inderscience Publishers, v. 3, n. 1, p. 4–21, 2011.

NIU, X.; WANG, L.; YANG, X. A comparison study of credit card fraud detection: Supervised versus unsupervised. *arXiv preprint arXiv:1904.10604*, 2019.

PADULA, A. J. A.; ALBUQUERQUE, P. H. M. Government corruption on brazilian capital markets: A study on lava jato (car wash) investigation. *Revista de Administração de Empresas*, Fundacao Getulio Vargas, v. 58, n. 4, p. 405–417, 2018.

PATEL, S.; GOND, S. Supervised machine (svm) learning for credit card fraud detection. *International Journal of Engineering Trends and Technology*, v. 8, p. 137–139, 2014.

PEROLS, J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, American Accounting Association, v. 30, n. 2, p. 19–50, 2011.

PHAM, T.; LEE, S. Anomaly detection in bitcoin network using unsupervised learning methods. *arXiv preprint arXiv:1611.03941*, 2016.

PHUA, C.; LEE, V.; SMITH, K.; GAYLER, R. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

PREVIC. 2020. Accessed: 28/02/2020. Disponível em: <http://www.previc.gov.br/>.

RAHIM, A. H. A.; RASHID, N. A.; NAYAN, A.; AHMAD, A.-R. Smote approach to imbalanced dataset in logistic regression analysis. **Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)**. [S.l.], 2019. p. 429–433.

RAJ, V.; MAGG, S.; WERMTER, S. Towards effective classification of imbalanced data with convolutional neural networks. **IAPR Workshop on Artificial Neural Networks in Pattern Recognition**. [S.l.], 2016. p. 150–162.

RAZA, M.; QAYYUM, U. Classical and deep learning classifiers for anomaly detection. **2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)**. [S.l.], 2019. p. 614–618.

ROY, A.; SUN, J.; MAHONEY, R.; ALONZI, L.; ADAMS, S.; BELING, P. Deep learning detecting fraud in credit card transactions. **2018 Systems and Information Engineering Design Symposium (SIEDS)**. [S.l.], 2018. p. 129–134.

SARKAR, S.; PRAMANIK, A.; MAITI, J.; RENIERS, G. Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. *Safety Science*, Elsevier, v. 125, p. 104616, 2020.

SHARPE, W. F. *Portfolio theory and capital markets*. [S.l.]: McGraw-Hill College, 1970.

SILVEIRA, D.; VASCONCELOS, S.; RESENDE, M.; CAJUEIRO, D. O. Won't get fooled again: A supervised machine learning approach for screening gasoline cartels. CESifo Working Paper Series, 2021.

SMITH, M. R.; MARTINEZ, T.; GIRAUD-CARRIER, C. An instance level analysis of data complexity. *Machine learning*, Springer, v. 95, n. 2, p. 225–256, 2014.

SOMAN, K.; LOGANATHAN, R.; AJAY, V. *Machine learning with SVM and other kernel methods*. [S.l.]: PHI Learning Pvt. Ltd., 2009.

TAHA, A. A.; MALEBARY, S. J. An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access*, IEEE, v. 8, p. 25579–25587, 2020.

TOMEK, I. et al. An experiment with the edited nearest-nieghbor rule. 1976.

TORGO, L.; BRANCO, P.; RIBEIRO, R. P.; PFAHRINGER, B. Resampling strategies for regression. *Expert Systems*, Wiley Online Library, v. 32, n. 3, p. 465–476, 2015.

VERDIKHA, N. A.; ADJI, T. B.; PERMANASARI, A. E. Study of undersampling method: Instance hardness threshold with various estimators for hate speech classification. *IJITEE (International Journal of Information Technology and Electrical Engineering)*, v. 2, n. 2, p. 39–44, 2018.

WANG, H.-Y. Combination approach of smote and biased-svm for imbalanced datasets. **2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)**. [S.l.], 2008. p. 228–231.

WANG, J.-H.; LIAO, Y.-L.; TSAI, T.-m.; HUNG, G. Technology-based financial frauds in taiwan: issues and approaches. **2006 IEEE International Conference on Systems, Man and Cybernetics**. [S.l.], 2006. v. 2, p. 1120–1124.

WANG, S.-C. Artificial neural network. In: *Interdisciplinary computing in java programming*. [S.l.]: Springer, 2003. p. 81–100.

WANG, W.; KAO, C.-c.; WANG, C. A simple model for detection of rare sound events. *arXiv preprint arXiv:1808.06676*, 2018.

WEISS, G. M. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, ACM, v. 6, n. 1, p. 7–19, 2004.

WEST, J.; BHATTACHARYA, M. Intelligent financial fraud detection: a comprehensive review. *Computers & security*, Elsevier, v. 57, p. 47–66, 2016.

WILSON, D. L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, IEEE, n. 3, p. 408–421, 1972.

ZHANG, H.; LI, Z.; SHAHRIAR, H.; TAO, L.; BHATTACHARYA, P.; QIAN, Y. Improving prediction accuracy for logistic regression on imbalanced datasets. **2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)**. [S.l.], 2019. v. 1, p. 918–919.

ZHANG, Y.; TONG, J.; WANG, Z.; GAO, F. Customer transaction fraud detection using xgboost model. **2020 International Conference on Computer Engineering and Application (ICCEA)**. [S.l.], 2020. p. 554–558.

ZHANG, Y.-P.; ZHANG, L.-N.; WANG, Y.-C. Cluster-based majority under-sampling approaches for class imbalance learning. **2010 2nd IEEE International Conference on Information and Financial Engineering**. [S.l.], 2010. p. 400–404.

ZHANG, Z.; ZHOU, X.; ZHANG, X.; WANG, L.; WANG, P. A model based on convolutional neural network for online transaction fraud detection. *Security and Communication Networks*, Hindawi, v. 2018, 2018.