CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

AVALIAÇÃO DE COMBINAÇÕES DE TÉCNICAS DE BALANCEAMENTO DE DADOS E DE SELEÇÃO DE ATRIBUTOS NA DETECÇÃO DE FRAUDE EM TRANSAÇÕES ONLINE DE CARTÃO DE CRÉDITO

Arthur Silveira

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Rafaelli de Carvalho Coutinho, D.Sc.
Eduardo Bezerra da Silva, D.Sc.

Rio de Janeiro,

Novembro de 2025

Avaliação de Combinações de Técnicas de Balanceamento de Dados e de Seleção de Atributos na Detecção de Fraude em Transações Online de Cartão de Crédito

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Arthur Silveira

Aprovada por:

Presidente, Profa. Rafaelli de Carvalho Coutinho, D.Sc. (orientadora)

Prof. Eduardo Bezerra da Silva, D.Sc. (coorientador)

Prof. Eduardo Ogasawara, D.Sc.

Prof. Ronaldo Ribeiro Goldschmidt, D.Sc.

Rio de Janeiro,

Novembro de 2025

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

S587 Silveira, Arthur
Avaliação de combinações de técnicas de balanceamento de dados e de seleção de atributos na detecção de fraude em transações online de cartão de crédito / Arthur Silveira. — 2025.
59f. : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2025.
Bibliografia : f. 51-59
Orientador: Rafaelli de Carvalho Coutinho
Coorientador: Eduardo Bezerra da Silva

1. Aprendizado de Máquina. 2. Fraude no cartão de crédito - Brasil. 3. Amostragem (Estatística). I. Coutinho, Rafaelli de Carvalho. (Orient.). II. Silva, Eduardo Bezerra da. (Coorient.). III. Título.

CDD 006.31

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

RESUMO

Avaliação de Combinações de Técnicas de Balanceamento de Dados e de Seleção de Atributos na Detecção de Fraude em Transações Online de Cartão de Crédito

Arthur Silveira

Orientadores:
Rafaelli de Carvalho Coutinho, D.Sc.
Eduardo Bezerra da Silva, D.Sc.

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

Devido ao aumento do comércio eletrônico e do uso de cartões de crédito, as fraudes com cartões de crédito tornaram-se um desafio para titulares de cartão, bancos emissores, bandeiras e lojistas. Apesar dos prejuízos, essas fraudes ainda representam uma pequena parte das transações, caracterizando um problema de desbalanceamento de dados nas áreas de detecção de fraudes do sistema financeiro. Este trabalho avalia várias combinações de técnicas de seleção de atributos, balanceamento de classes e algoritmos de classificação. Para balancear as classes, foram usadas técnicas de subamostragem, superamostragem e ajustes de limiares nos classificadores. O objetivo é criar um pipeline que permita a reprodutibilidade e ampliação das avaliações dessas técnicas. As combinações foram testadas em dois conjuntos de dados desbalanceados, avaliados pela métrica escore F_1. Os resultados mostram um ganho de desempenho quando são implementadas técnicas de balanceamento de dados e otimização de limiares de classificação.

Palavras-chave:
Aprendizado de Máquina, Classificação, Dados Desbalanceados, Desbalanceamento de Classes, Detecção de Fraudes, Fraude em Cartão de Crédito, Otimização de Limiares de Saída, Sobreamostragem, Subamostragem, Técnicas de Amostragem

Rio de Janeiro,

Novembro de 2025

ABSTRACT

Avaliação de Combinações de Técnicas de Balanceamento de Dados e de Seleção de Atributos na Detecção de Fraude em Transações Online de Cartão de Crédito

Arthur Silveira

Advisors:
Rafaelli de Carvalho Coutinho, D.Sc.
Eduardo Bezerra da Silva, D.Sc.

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Due to the increase in e-commerce and credit card use, fraud detection has become a challenge for cardholders, issuing banks, brands and retailers. Despite the losses, this type of fraud still represents a small portion of transactions, defining a data imbalance problem in the areas of fraud detection within the financial system. This work evaluates various combinations of feature selection techniques, class balancing, and classification. Undersampling, oversampling techniques, and classifier threshold adjustments were used to balance the classes. The objective is to create a pipeline that allows the reproducibility and expansion of evaluations of these techniques. The combinations were tested on two imbalanced datasets and evaluated using the F_1-$score$ metric. The results show improved performance when implementing data balancing techniques and optimal classification threshold adjustments.

Key-words:
Class Imbalance, Classification, Credit Card Fraud, Fraud Detection, Imbalanced Data, Machine Learning, Output Thresholding, Oversampling, Sampling Techniques, Undersampling

Rio de Janeiro,

Novembro de 2025

Sumário

I Introdução **1**

II Referencial Teórico **4**

II.1 Visão Geral de Fraudes de Cartão de Crédito 4

II.2 Engenharia de Atributos 5

II.3 Amostragem com Estratificação 6

II.4 Técnicas de Balanceamento de Classes 7

II.4.1 Superamostragem 7

II.4.2 Subamostragem 8

II.5 Técnicas de Detecção de Fraudes de Cartão de Crédito 9

II.5.1 Aprendizado Não-Supervisionado 9

II.5.2 Aprendizado Supervisionado 10

II.6 Métricas de Desempenho 13

II.7 Validação Cruzada 14

II.8 Busca por Limiar Ótimo 15

III Trabalhos Relacionados **17**

III.1 Busca Sistemática 17

III.2 Organização do Conjunto de Dados 18

III.3 Utilização de Técnicas de Seleção de Atributos 19

III.4 Abordagem para as Técnicas de Balanceamento de Classes 19

III.4.1 Proposta de Novas Técnicas de Balanceamento com Comparativo 19

III.4.2 Proposta de Novas Técnicas de Balanceamento sem Comparativo 21

III.4.3 Comparativo entre Técnicas de Balanceamento 22

III.4.4 Aplicação de Técnicas de Balanceamento 22

III.5 Técnicas de Estratificação de Conjuntos de Dados e Validação Cruzada 23

III.6 Implementação de Múltiplos Classificadores 24

III.7 Métricas de Desempenho apropriadas para Desbalanceamento 25

III.8 Síntese dos Trabalhos 25

IV Metodologia **29**

IV.1 Conjuntos de Dados 29

IV.2 Pré-Processamento 30

IV.2.1 Tratamento dos Dados 31

IV.2.2 Divisão 32

IV.2.3 Balanceamento de Classes 32

IV.2.4 Seleção de Atributos 33

IV.3 Treinamento do Classificador 34

IV.4 Avaliação 34

V Resultados **35**

V.1 Configuração dos Experimentos 35

V.2 Análise Exploratória da Base IEEE-CIS 35

V.3 Análise de Experimentos 42

V.3.1 Análise de Técnicas de Seleção de Atributos, de Balanceamento de Classes e Classificadores 42

V.3.2 Análise com Codificação *One-Hot* 44

V.3.3 Análise com Limiares Ótimos, Estratificação de Dados e Validação Cruzada 44

V.4 Resultados Alcançados 47

VI Conclusões **49**

Referências Bibliográficas 51

Lista de Figuras

II.1 Modelo de operações diárias de um departamento anti-fraude. 6

II.2 Lógica da validação cruzada estratificada. 15

III.1 Mapeamento sistemático utilizado no trabalho. 18

IV.1 Visão geral da metodologia definida para os experimentos. 29

V.1 Quantidade de transações por tipo de produto e variável-alvo. 36

V.2 Boxplots dos valores por transação (em USD), por tipo de Produto e variável-alvo. 37

V.3 Quantidade de transações por bandeira e variável-alvo. 37

V.4 Boxplots dos valores por transação (em USD), por tipo de produto e variável-alvo. 38

V.5 Quantidade de transações por tipo de pagamento e variável-alvo. 38

V.6 Boxplots dos valores por transação (em USD), por tipo de pagamento e variável-alvo. 39

V.7 Total de transações por dia no mês e percentual de fraude. 39

V.8 Total de transações (em milhões de USD) por dia no mês e percentual de fraude. 40

V.9 Total de transações por dia da semana no mês e percentual de fraude. 40

V.10 Total de transações (em bilhões de USD) por dia da semana no mês e percentual de fraude. 41

V.11 Total de transações por hora no mês e percentual de fraude. 41

V.12 Total de transações (em milhões de USD) por hora no mês e percentual de fraude. 41

Lista de Tabelas

II.1 Matriz de confusão 13

III.1 Comparativo dos Trabalhos Relacionados 27
III.2 Comparativo dos Trabalhos Relacionados (cont.) 28
III.3 Resumo Quantitativo dos Trabalhos Relacionados 28

IV.1 Lista final dos atributos selecionados para treinamento (conjunto de dados utilizado: IEEE-CIS) 31

V.1 Desempenho dos modelos (via escore F_1) por configuração experimental (conjunto de dados utilizado: IEEE-CIS) 43
V.2 Desempenho dos modelos (via escore F_1) por configuração experimental, após codificação *one-hot* (conjunto de dados utilizado: IEEE-CIS) 45
V.3 Desempenho dos modelos (via escore F_1) por configuração experimental, após implementação de limiares ótimos, estratificação de dados e validação cruzada (conjunto de dados utilizado: IEEE-CIS) 46
V.4 Síntese dos resultados que alcançaram melhores desempenhos de escore F_1 48

Lista de Abreviações

AB	*AdaBoost*	11, 21, 23, 34, 44
AC	Análise De Correlação	6, 42
AD	Árvore De Decisão	10, 11, 12, 25, 34, 35, 42, 44, 47, 48, 49
ADASYN	*Adaptive Synthetic Sampling*	8, 22, 25, 33, 42, 44, 47
AED	*Análise Exploratória de Dados*	35
ANOVA	Análise De Variância	6, 26, 42
AUC	*Area Under the ROC Curve*	14, 25, 28
AUPRC	*Area Under the Precision-Recall Curve*	14
BSMOTE	*Borderline Synthetic Minority Oversampling Technique*	8, 32, 42, 44, 47, 48, 49
CB	*CatBoost*	11, 21, 34, 44
DBSMOTE	*Density-Based Synthetic Minority Over-Sampling Technique*	22
ENN	*Edited Nearest Neighbours*	21, 22
ET	*Extremely Randomized Trees*	10, 11, 25, 34, 35, 42, 44, 47, 48, 49
FA	Floresta Aleatória	11, 21, 23, 25, 34, 42, 44
FN	Falsos Negativos	13
FP	Falsos Positivos	13, 14
GAN	*Generative Adversarial Network*	19, 20, 21, 22, 50
GB	*Gradient Boosting*	11, 21, 25, 34, 42, 44
GMUS	*Gaussian Mixture Undersampling*	20
IEEE-CIS	*Institute of Electrical and Electronics Engineers - Computational Intelligence Society*	29, 30, 31, 32, 33, 34, 42, 47
IHT	*Instance Hardness Threshold*	9, 33, 42, 47, 48
KNN	*K-Nearest Neighbors*	12, 23, 34, 42, 44
LGBM	*Light Gradient-Boosting Machine*	21
LLM	*Large Language Model*	1
MLG	*Machine Learning Group*	30, 31, 32, 33, 34, 47, 48
NB	*Naive Bayes*	12, 23, 34, 42, 44
NSMOTE	*Navo Minority Oversampling Technique*	20
QQ	Qui-Quadrado	6, 26, 33, 34, 42, 44, 47, 49

RL	Regressão Logística	12, 13, 23, 34, 42, 44
ROC	*Receiver Operating Characteristic*	14
ROS	*Random Oversampling*	7, 32, 42, 44, 47
RUS	*Random Undersampling*	8, 19, 20, 21, 22, 23, 25, 32, 42, 47
SBBD	Simpósio Brasileiro De Banco De Dados	49
SDF	Sistemas De Detecção De Fraude	1, 5, 7
SMOTE	*Synthetic Minority Oversampling Technique*	8, 19, 20, 21, 22, 23, 25, 32, 42, 44, 47, 49
SMOTE-TL	*Synthetic Minority Oversampling Technique-Tomek Links*	22
SMOTE-ENN	*Synthetic Minority Oversampling Technique-Edited Nearest Neighbours*	21, 22
SPF	Sistemas De Prevenção À Fraude	4
SVM	*Support Vector Machine*	23
TL	*Tomek Links*	22
ULB	*Université Libre de Bruxelles*	30, 31, 32, 33, 34, 47, 48
VN	Verdadeiros Negativos	13, 14
VP	Verdadeiros Positivos	13, 14
XGBOOST	*Extreme Gradient Boosting*	19, 21, 22, 23, 25

Capítulo I Introdução

Com o amplo acesso à internet no dia a dia e, mais recentemente, com a pandemia de coronavírus em 2020, pessoas têm utilizado cada vez mais a navegação *on-line* para fazer compras por meio de seus cartões de crédito [Smiles and Kamalakannan, 2020][1]. Apesar das cifras desse mercado crescerem ano após ano, crescem também oportunidades para atividades fraudulentas a partir do roubo de dados de cartões, as quais trazem prejuízos para compradores, lojas, bancos e empresas de cartão de crédito [Rodrigues et al., 2022]. Segundo o *Global Fraud Report*[2] de 2024 da empresa Cybersource[3], globalmente, o número médio de diferentes tipos de ataques de fraude que comerciantes experimentaram de 2023 para 2024 aumentou. Além disso, cerca de metade dos comerciantes planeja aumentar os gastos em ferramentas/tecnologias (60%), com um consenso de que as áreas de maior foco para melhoria no próximo ano incluem ferramentas de gerenciamento de fraude baseadas em inteligência artificial e aprendizado de máquinas.

Para varejistas e instituições financeiras em geral, a detecção de fraudes em transações *on-line* é desafiadora porque os métodos de fraude evoluem e estão em processo de aperfeiçoamento pelos criminosos [Correa Bahnsen et al., 2016; Porwal and Mukund, 2019]. Os sistemas de detecção devem, idealmente, identificar esse comportamento imediatamente e bloquear a transação, a fim de preservar a integridade da cadeia [Artikis et al., 2017]. Além disso, a necessidade da criação de Sistemas de Detecção de Fraude (SDF) precisos e automatizados é fundamental, dado o volume de compras feitos diariamente. Para tal, algoritmos de aprendizado de máquina, como classificadores e algoritmos de clusterização, são estudados e utilizados [Bhattacharyya et al., 2011; Zhang et al., 2020]. No entanto, mesmo com a criação de um SDF que utilize algoritmos de aprendizado de máquina para dar escala ao processo de identificação de fraudes, outras questões ainda surgem com relação à construção desses modelos.

Uma dessas questões é o fato de transações fraudulentas corresponderem a uma parte pequena do volume que é processado diariamente, caracterizando o chamado desbalanceamento de classes. Sob o ponto de vista da mineração de dados, isso gera dificuldades em áreas chamadas sensíveis a custo,

[1] Declaração de Uso de Ferramenta de Escrita: O autor declara que a assistência de modelos de *Large Language Model* (LLM) foi utilizada exclusivamente para aprimoramento gramatical, estilístico e para garantir a clareza textual deste manuscrito. LLM não teve qualquer envolvimento na formulação da metodologia, análise de dados, resultados experimentais ou conclusões da pesquisa.

[2] `https://www.visaacceptance.com/content/dam/documents/campaign/fraud-report/global-fraud-report-2024.pdf`

[3] Cybersource é uma empresa especializada em prevenção à fraude e uma das referências no assunto.

como fraudes de cartão de crédito [Correa Bahnsen et al., 2016; Muaz et al., 2020]. Nessa área, um falso positivo (isto é, uma transação genuína erroneamente classificada como fraudulenta) acarreta prejuízos menores do que um falso negativo (nesse caso, uma transação fraudulenta classificada como genuína): danos severos ocorrem se a classe minoritária (transações fraudulentas) for classificada como majoritária (transações genuínas), enquanto o contrário compromete menos as instituições envolvidas [Ali et al., 2015]. Assim, em conjuntos de dados desbalanceados, é comum que a classe minoritária seja ignorada durante o processo de aprendizado e acabe sendo tratada, inclusive, como ruído, de acordo com o trabalho seminal para dados desbalanceados de He and Garcia [2009].

Consequentemente, motivações para pesquisas de aprendizado em dados desbalanceados envolvem as mais variadas soluções para diminuir essa diferença entre a classe majoritária e a minoritária [SUN et al., 2009]. Abordagens para além de técnicas de balanceamento de dados costumam ser utilizadas, como a aprendizagem sensível a custo, que incorpora os custos de diferentes tipos de erros (falsos positivos e falsos negativos) diretamente no processo de treinamento do modelo [He and Garcia, 2009]. Além dessa, existem métodos de aprendizado baseados em *kernel*, que transformam os dados em um espaço de alta dimensionalidade, em que é mais fácil separar as classes, e o aprendizado ativo, onde o algoritmo seleciona de forma iterativa os exemplos de treinamento mais informativos para serem rotulados por um especialista [He and Garcia, 2009].

No entanto, a forma mais comum de aprendizado é, de fato, por meio de técnicas de balanceamento de dados [He and Garcia, 2009]. Também chamadas de abordagens em nível de dados, em que existe a manipulação direta dos dados antes de aplicar algoritmos de aprendizado, essas técnicas envolvem métodos de subamostragem, superamostragem e até métodos híbridos [Kotsiantis et al., 2005]. Diversos estudos demonstram que um conjunto de dados balanceado oferece melhor desempenho, o que reforça sua utilização em contextos de alto desbalanceamento [He and Garcia, 2009]. Os já citados métodos de aprendizado ativo e baseados em *kernel* requerem maior complexidade computacional e uma quantidade significativa de rotulação manual, o que pode não ser viável em alguns cenários [He and Garcia, 2009]. Também já citados, métodos sensíveis a custo dependem fortemente de uma precisa e, por vezes, subjetiva atribuição de custos, que pode variar entre diferentes aplicações. Portanto, técnicas de balanceamento de dados oferecem uma abordagem direta e aplicável para melhorar a detecção de fraudes, potencializando o desempenho dos classificadores sem a necessidade de ajustes complexos ou conhecimento especializado adicional.

Apesar do aumento no uso de técnicas de aprendizado de máquina para detectar fraudes em transações com cartões de crédito, a literatura ainda apresenta lacunas quanto à análise combinada e integrada de múltiplas técnicas de balanceamento de classes, métodos de seleção de atributos e classificadores. Além disso, a generalização dos métodos é limitada, pois muitos estudos validam suas propostas em um único conjunto de dados. Além disso, observa-se uma carência na análise de diferentes combinações

de múltiplas técnicas de balanceamento de classes com diferentes classificadores e métodos de seleção de atributos. A ausência de uma abordagem integrada e comparativa que envolva múltiplos cenários e métricas específicas para dados desbalanceados evidencia um espaço ainda pouco explorado e de relevância prática.

Dessa forma, esse trabalho busca preencher essas lacunas, respondendo à pergunta de pesquisa sob quais combinações para tratamento, balanceamento, seleção de atributos e classificadores, validadas em mais de um conjunto de dados real e desbalanceado, essas combinações tem melhor desempenho e por quê. A análise foca na identificação dos cenários em que tais combinações oferecem melhor desempenho na detecção de fraudes, especialmente sob a ótica da classe minoritária. Essas combinações, que também são chamadas nesse trabalho de *pipeline* de dados, envolvem outras etapas de preparação (como, por exemplo, limpeza de dados) para que os modelos de aprendizado de máquina possam ser treinados da forma mais eficaz possível. Vale ressaltar que, para o uso do *pipeline* desenvolvido e sua avaliação, foram utilizados dois conjuntos de dados de transações de cartões de crédito disponíveis publica e altamente desbalanceados, o que também exigiu, dentro desses cenários, a utilização de estratificação de dados de treino e teste e sua validação cruzada durante o treinamento dos classificadores.

Assim, a principal contribuição desse trabalho é a apresentação de um comparativo que evidencia os cenários em que há ou não melhora significativa de desempenho, a partir da análise de combinações de técnicas de balanceamento, de seleção de atributos e classificadores. Isso decorre principalmente do fato de que o *pipeline* de dados implementado também é abrangente para o tratamento, balanceamento e classificação de transações de cartão de crédito como fraudulentas ou não. Além disso, a própria validação desse *pipeline* ocorreu em mais de um conjunto de dados e focou em métricas de desempenho apropriadas para conjuntos desbalanceados.

O restante deste trabalho é organizado como segue. O Capítulo II descreve conceitos fundamentais para a compreensão deste trabalho. O Capítulo III apresenta um levantamento dos artigos relacionados ao tema. O Capítulo IV traz detalhes sobre a metodologia utilizada neste trabalho, apresentando as técnicas de balanceamento de dados, os algoritmos de aprendizado de máquina e sua avaliação. O Capítulo V mostra os resultados experimentais desse processo. Por fim, o Capítulo VI discute esses resultados e aponta trabalhos futuros.

Capítulo II Referencial Teórico

Neste capítulo, são apresentados os conceitos fundamentais relacionados à detecção de fraudes de cartão de crédito em conjuntos de dados desbalanceados. Os conceitos envolvem uma visão geral de fraudes de cartão de crédito (Seção II.1), engenharia de atributos (Seção II.2), amostragem com estratificação (Seção II.3), técnicas de balanceamento de classes (Seção II.4), técnicas de detecção de fraudes de cartão de crédito (Seção II.5), métricas de desempenho (Seção II.6), validação cruzada (Seção II.7) e busca por limiar ótimo (Seção II.8).

II.1 Visão Geral de Fraudes de Cartão de Crédito

Uma fraude de cartão de crédito se caracteriza como uma transação feita por uma pessoa não autorizada enquanto titulares de cartão, bancos emissores, bandeiras e lojistas não estão cientes da existência dessa transação no momento da sua realização [Askari and Hussain, 2017]. A fraude costuma ser de dois tipos: operacional e comportamental [Bhattacharyya et al., 2011]. No primeiro caso, fraudadores obtem novos cartões junto aos emissores utilizando informações falsas ou de outras pessoas; já no segundo caso, existem quatro subtipos principais: (i) por extravio do cartão físico: quando fraudadores interceptam cartões de crédito enviados pelos emissores antes que cheguem aos seus titulares; (ii) por roubo/perda do cartão físico; (iii) por contrabando *online* de informações: como, por exemplo, por *phishing* ou compra na *deep web*; e (iv) por cartão ausente (do inglês, *Cardholder Not Present (CNP)*), em que informações são utilizadas remotamente para compras online ou por outros meios [Abdallah et al., 2016].

Mecanismos de prevenção e detecção de fraude são as formas mais comuns de se combater esse tipo de ação e, apesar de parecerem sinônimos, funcionam de diferentes formas [Abdallah et al., 2016; Rodrigues et al., 2022]. Sistemas de Prevenção à Fraude (SPF) servem para bloquearem a fraude desde o seu primeiro momento, por meio de de mecanismos de restrição, eliminação ou destruição de ataques cibernéticos a *hardwares*, *softwares*, redes e/ou dados das plataformas de compras online [Abdallah et al., 2016]. Exemplos desses mecanismos vão desde algoritmos de criptografia a *firewalls*. No entanto, mesmo com a evolução desses mecanismos de prevenção, fraudadores ainda conseguem atravessar essas restrições e continuar com a ação.

Nesse momento, mecanismos de detecção de fraude entram em ação, buscando identificar atividade-

des anômalas e, dependendo do fluxo de regras configurado no sistema, bloqueá-las ou reportá-las a analistas [Abdallah et al., 2016]. Dado o volume atual das transações em plataformas de compras online, SDF automatizados ganham importância com a evolução da área de mineração de dados e que oferece vantagens como identificação automática de padrões de comportamento fraudulento já conhecidos e a revelação de novos padrões.

A Figura II.1 reflete um modelo de operações diárias de um departamento anti-fraude em uma empresa de comércio eletrônico [Tax et al., 2021]:

- plataforma de *e-commerce*: serviços de compra e venda online, como Amazon, Magazine Luiza, Mercado Livre, Shein e Shopee;
- usuários: divididos entre genuínos e fraudulentos, eles interagem com a plataforma (1), que, por sua vez, gera os dados (2);
- dados: sob a ótica do aprendizado de máquina, são gerados sob a forma de variáveis e rótulos. As variáveis apresentam comportamento relevante, como dados da compra e do perfil do usuário, data e horário da transação. Rótulos geralmente resultam do trabalho dos investigadores (10) e da notificação de bancos emissores, bandeiras, adquirentes e processadoras de pagamento responsáveis pela autorização e liquidação (12);
- investigadores: colaboradores da empresa de comércio eletrônico que recebem os casos suspeitos de fraude (3, 9, 11) a partir de regras definidas no sistema de decisão (5). Ações manuais podem ser feitas, como bloqueio da compra e/ou do usuário (4) e ajustes nas regras do sistema (5);
- sistema de decisão: sistema automatizado que toma ações concretas sobre as transações, seja diretamente sobre as compras ou usuários (8), ou indiretamente, indicando investigações (9). Essas ações resultam de combinações de regras formuladas pelos investigadores (5) ou a partir do treinamento de modelos de aprendizado de máquina (7);
- modelo: treinamento e teste dos dados usando os modelos de aprendizado de máquina (6).

II.2 Engenharia de Atributos

Os conjuntos de dados de fraude com dimensão alta (*i.e.*, muitas colunas, atributos ou variáveis) tornam o processo de mineração de dados mais complexo e o de detecção das fraudes, mais lento [Abdallah et al., 2016]. Isso decorre da chamada "maldição da dimensionalidade", em que, quanto maior a dimensão de um conjunto de dados, maior a demanda pelo seu armazenamento e processamento - e, consequentemente, pior o desempenho computacional do sistema [Jia et al., 2022]. É por isso que, nesse contexto, SDF implementados atualmente utilizam abordagens de redução de dimensionalidade,

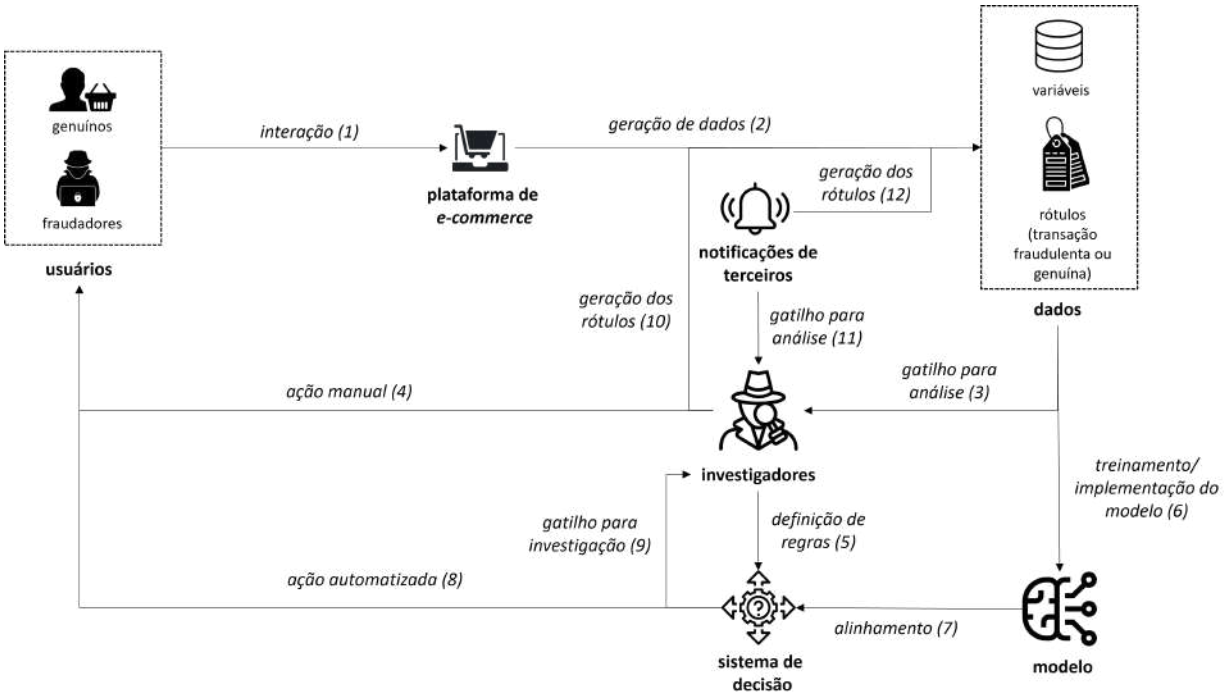


Figura II.1: Um modelo de operações diárias de um departamento anti-fraude em uma empresa de comércio eletrônico (adaptado de [Tax et al., 2021])

como técnicas de seleção de atributos, em que os atributos mais relevantes são selecionados por meio de técnicas como Análise de Correlação (AC) e Análise de Variância (ANOVA) - cujo método de seleção considera relações não-lineares com a variável-alvo - e teste do Qui-Quadrado (QQ) - que, por outro lado, considera relações lineares - para a construção do modelo.

Abordagens de redução de dimensionalidade ganham ainda mais importância quando são utilizadas técnicas como a codificação *one-hot* (do inglês, *one-hot encoding*). Essa técnica busca representar cada variável categórica por um vetor binário, no qual apenas um dos elementos é 1 (indicando a presença da categoria) e os outros são 0 (indicando a ausência) [Jahnavi et al., 2023]. Essa técnica é útil quando as categorias não têm uma ordem específica e quando se deseja evitar a atribuição de relações de ordem ou magnitude entre elas.

II.3 Amostragem com Estratificação

A amostragem com estratificação é aplicada para garantir que o número de amostras para treinamento de cada classe está balanceada, permitindo que a estrutura original dos dados seja preservada [Sadaiyandi et al., 2023]. O procedimento padrão é selecionar aleatoriamente exemplos de ambas as classes e depois combiná-los em conjuntos de dados para o treinamento do classificador. Como existe uma tentativa de garantir o balanceamento das classes para o treinamento, esse método revela-se bastante útil para conjuntos de dados desbalanceados [Kathiravan et al., 2023].

II.4 Técnicas de Balanceamento de Classes

O desbalanceamento de classes é considerado um dos fatores mais críticos relacionado a SDF, principalmente por conta do seu caráter intrínseco - isto é, como resultado da baixa frequência natural de fraudes em comparação às transações legítimas [Abdallah et al., 2016; He and Garcia, 2009]. Por isso, técnicas de balanceamento de classes são cruciais para diminuir diferenças severas na proporção de classes em um conjunto de dados. Caso não sejam implementadas, o desbalanceamento pode prejudicar o desempenho dos algoritmos de aprendizado de máquina, aumentando as chances de falsos positivos e negativos e todos os impactos financeiros decorrentes disso [Muaz et al., 2020].

Essas técnicas podem ser divididas entre três grandes grupos: técnicas a nível de dados, técnicas a nível de algoritmo e técnicas a nível de sensibilidade a custo [SUN et al., 2009]. A nível de dados, elas são implementadas como parte do pré-processamento do conjunto de dados - isto é, antes de sua aplicação em algoritmos de aprendizado de máquinas -, justamente para equilibrar a representatividade das classes e remover ruídos [Abdallah et al., 2016]. A nível de algoritmo, os ajustes são feitos no momento do treinamento dos algoritmos, seja por meio dos parâmetros dos algoritmos utilizados ou utilizando algoritmos mais apropriados para situações de desbalanceamento, podendo apresentar uma resistência maior a erros de classificação nesse tipo de situação [SUN et al., 2009]. Finalmente, a nível de custo, um peso maior a falsos positivos e negativos criados ao longo do processo de treinamento é aplicado, por meio de uma matriz de custo que pondera esses resultados [SUN et al., 2009].

Por esse motivo, as técnicas a nível de dados são melhor exploradas nesta pesquisa. Dentro delas, as principais categorias são superamostragem (do inglês, *oversampling*) e subamostragem (do inglês, *undersampling*) [Alfaiz and Fati, 2022].

II.4.1 Superamostragem

De forma geral, na superamostragem, busca-se gerar novas entradas no conjunto de dados da classe minoritária e aumentar sua proporção em relação ao conjunto de dados como um todo. As técnicas implementadas variaram de replicações das classes minoritárias até geração sintética de novos valores, baseados no conjunto existente.

No caso das replicações, duas técnicas merecem ser mencionadas: a primeira, *resample* é um método de replicação simples e aleatório de dados da classe minoritária, mas com um processo de implementação manual[1]. Adicionalmente, é preciso separar os registros da classe minoritária - e a própria classe minoritária em si - do restante do conjunto com a classe majoritária, aplicar o método e devolver o conjunto tratado ao original. A segunda, *Random Oversampling* (ROS), é também um método aleatório de replicação de dados classe minoritária, mas sem a necessidade de separação dela do restante do

[1] https://scikit-learn.org/stable/modules/generated/sklearn.utils.resample.html

conjunto de dados [He and Garcia, 2009]. Nos dois casos, essa replicação pode levar a uma situação de sobreajuste (do inglês, *overfitting*): o desempenho do algoritmo, durante o treinamento, pode ser muito satisfatório; no entanto, ao mesmo tempo, perde-se a capacidade de generalização e um desempenho pior é verificado na fase de testes [Galar et al., 2012; Kotsiantis et al., 2005; He and Garcia, 2009].

No caso da geração sintética, a principal técnica utilizada é *Synthetic Minority Oversampling Technique* (SMOTE), a qual cria artificialmente novos dados nos espaços entre as entradas já existentes da classe minoritária [He and Garcia, 2009]. Isso ocorre por meio da seleção aleatória de uma entrada e da interpolação de k instâncias próximas para a criação do novo dado, o que permite um risco menor de sobreajuste [Galar et al., 2012; Kotsiantis et al., 2005]. SMOTE tem se mostrado um método bem sucedido nas suas implementações, mostrando desempenho melhor do que uma simples replicação de dados já existentes da classe minoritária [Vimalraj and Dr.R, 2018].

Variações de SMOTE também foram implementadas. No caso da *Borderline Synthetic Minority Oversampling Technique* (BSMOTE), prioriza-se a geração de novos dados a partir de instâncias da classe minoritária chamadas de limítrofes (do inglês, *borderline*) [Prati et al., 2009]. Normalmente, essas instâncias são ambíguas e podem tanto ser classificadas como fraudulentas quanto genuínas. Assim, o classificador fortalece essa região com mais exemplos e seu desempenho melhora [Alfaiz and Fati, 2022].

No entanto, a SMOTE pode criar, em algumas situações, novas instâncias que acabam sendo muito semelhantes entre si ou até mesmo às entradas originais do conjunto de dados. A técnica *Adaptive Synthetic Sampling* (ADASYN) corrige isso, que também pode ser interpretada como uma extensão do próprio SMOTE [He and Garcia, 2009]. A ideia principal do algoritmo é, iterativemente, modificar os pesos dados à geração das novas instâncias para compensar uma assimetria na distribuição desses novos dados.

II.4.2 Subamostragem

Na subamostragem, busca-se reduzir a quantidade de entradas da classe majoritária com a intenção de diminuir sua proporção, funcionando de forma inversa à superamostragem. A técnica *resample* proporciona um processo de exclusão aleatória de dados da classe majoritária, em que persiste o aspecto manual da implementação: como no caso de *resample* para superamostragem, também é preciso separar os atributos da classe majoritária, e a própria classe majoritária em si, do restante do conjunto, aplicar o método de eliminação e devolver o conjunto tratado ao original. Já no caso da *Random Undersampling* (RUS), trata-se também de um método aleatório de eliminação de dados - no caso, da classe majoritária -, ainda sem a necessidade de separação dessa classe do restante do conjunto de dados [Hasib et al., 2020]. No entanto, o impacto dessa remoção é diferente, já que pode afetar dados importantes para a classificação e, por consequência, o desempenho do algoritmo [He and Garcia, 2009; Kotsiantis et al., 2005; Kaur et al., 2019].

A técnica *Near Miss* funciona selecionando exemplos da classe majoritária com base na proximidade aos exemplos da classe minoritária [Prati et al., 2009]. Isso é feito por meio do cálculo da distância média das instâncias da classe majoritária com relação às instâncias da classe minoritária: quanto menor forem essas distâncias, maiores as chances dos exemplos da classe majoritária serem retidos, eliminando-se exemplos mais distantes dessa fronteira de decisão. A premissa, nesse caso, é a de que instâncias relativamente próximas contribuem mais para o aprendizado da fronteira de decisão do modelo e, consequentemente, no seu desempenho. Apesar disso, ao eliminar exemplos da classe majoritária que estão mais distantes, pode-se remover dados relevantes para a generalização do modelo.

Finalmente, a *Instance Hardness Threshold* (IHT), diferentemente dos métodos baseados em vizinhos mais próximos, considera a dificuldade de classificar cada instância da classe minoritária e baseia-se nela para realizar a subamostragem [Le et al., 2018]. O algoritmo conduz esse processo ao: (i) treinar um modelo inicial sobre o conjunto de dados desbalanceado; (ii) calcular a dificuldade de classificação das classes majoritárias e (iii) removê-las, considerando um limiar de dificuldade de classificação pré-estabelecido. Assim, o objetivo de IHT é manter as instâncias da classe majoritária que são facilmente classificáveis, enquanto remove aquelas que são difíceis de classificar.

II.5 Técnicas de Detecção de Fraudes de Cartão de Crédito

As técnicas de detecção de fraudes de cartão de crédito são normalmente classificadas em duas categorias gerais: aprendizado não-supervisionado e aprendizado supervisionado [Sorournejad et al., 2016]. No primeiro caso, o objetivo é detectar transações em um conjunto de dados em que elas não tenham sido previamente identificadas como fraudulentas ou não; já no segundo, são aplicadas técnicas de aprendizado em conjuntos de dados cujas transações já tenham sido previamente rotuladas como legítimas ou fraudulentas.

II.5.1 Aprendizado Não-Supervisionado

Dado que o contexto é de detectar transações em um conjunto de dados não rotulado, analisa-se a transação sob a perspectiva do comportamento usual de um usuário, sendo necessário compreender previamente as ações de um comprador para que atitudes pouco usuais - e, geralmente, fraudulentas - possam ser detectadas [Sorournejad et al., 2016]. Embora exija certo conhecimento prévio sobre o comportamento do usuário, a principal vantagem do aprendizado não supervisionado está no fato de que não há necessidade de uma rotulação prévia de cada transação do conjunto de dados - o que pode ser intensivo em homem-hora e custos - para treinamento, teste e implementação do modelo.

As principais técnicas de implementação de aprendizado não-supervisionado envolvem o agrupamento dos dados, cujos algoritmos permitem a identificação e separação de diferentes distribuições de dados [Gupta et al., 2022; Carcillo et al., 2021]. Essa separação pode ser feita de diferentes formas, prin-

cipalmente por meio de algoritmos baseados em: (i) densidade, em que medidas de similaridade criam regiões mais densas que outras; (ii) partição, em que os dados são aleatoriamente separados para depois serem aplicadas medidas de similaridade, caracterizando o agrupamento e (iii) hierarquia, em que se criam agrupamentos semelhantes a árvores, respeitando uma hierarquia subjacente aos dados [Gupta et al., 2022].

II.5.2 Aprendizado Supervisionado

Trata-se da forma mais comum de modelagem para detecção de fraudes de cartão de crédito [Abdallah et al., 2016]: a partir dessa rotulação prévia, o conjunto de dados é utilizado para a criação de modelos de classificação capazes de prever a legitimidade das novas entradas ou não. De forma geral, esse tipo de modelagem é provada como efetiva na identificação de novos comportamentos fraudulentos que já tenham sido vistos anteriormente [Sorournejad et al., 2016]. No entanto, também pode apresentar algumas limitações, como o próprio processo de rotulação, que muitas vezes é manual e dependente da ação dos investigadores da área de fraudes, e o risco de potenciais falsos positivos ou negativos, dada a falibilidade do processo de julgamento humano nesse tipo de situação [Abdallah et al., 2016].

Por tratar-se do mecanismo mais comum de detecção de fraudes, técnicas de aprendizado supervisionado serão analisadas com maior profundidade, sob o contexto de classificação binária (fraude ou não), característico desse tipo de situação. Abaixo, seguem algumas das técnicas implementadas.

Árvore de Decisão (AD)

AD é um dos algoritmos mais utilizados em problemas de classificação, especialmente por conta da sua capacidade de lidar com problemas complexos de forma compreensível e de fácil explicação [Tran and Dang, 2021]. É um método que possui nós internos que representam os atributos do conjunto de dados, conexões entre esses nós (chamados ramos) que representam os resultados desses atributos e folhas que produzem os rótulos necessários para a classificação das entradas do conjunto de dados [Alfaiz and Fati, 2022].

Existem duas grandes etapas nesse método: (i) a construção da árvore em si e (ii) processos iterativos para o reconhecimento dos padrões e classificação das entradas do conjunto de dados [Tran and Dang, 2021]. A construção da árvore é obtida ao subdividir o conjunto de dados, sendo baseada no resultado da análise dos atributos pelos nós internos e ramos, enquanto os processos iterativos repetem essas análises até obterem o rótulo final [Afriyie et al., 2023].

Extremely Randomized Trees (ET)

A popularidade de modelos de AD em mineração de dados surge da facilidade em sua implementação, flexibilidade ao lidar com diversos tipos de atributos e interpretação [Bhattacharyya et al., 2011].

No entanto, esses mesmos modelos podem ser instáveis e sensíveis a anomalias introduzidas durante o treinamento. Dessa forma, modelos *ensemble* procuram endereçar esse tipo de situação ao desenvolver diferentes modelos e agregá-los para a previsão da variável-alvo, como no caso de Floresta Aleatória (FA), que consiste em um conjunto de AD para classificação [Tran and Dang, 2021; Muaz et al., 2020].

O método *Extremely Randomized Trees* (ET), também chamado de *Extra Trees*, é conhecido como o classificador mais randômico da família de FA [J and Senthilselvi, 2022]. No entanto, ao contrário do método tradicional de FA, essa abordagem de aprendizado por comitê (do inglês *ensemble learning*) combina os resultados de diversas árvores de decisão baseado em divisões randômicas do conjunto de dados [Devarakonda et al., 2023].

Floresta Aleatória (FA)

FA é um método que agrega diversas árvores de decisão, podendo ser usado para construir modelos preditivos tanto para problemas de classificação quanto de regressão. Ele gera várias árvores de decisão, onde cada árvore individual atua como um aprendiz fraco, mas, juntas, formam um aprendiz com melhor desempenho [Alfaiz and Fati, 2022]. Em geral, uma FA é um método eficaz e rápido ao lidar com conjuntos de dados desbalanceados, mesmo quando esses conjuntos possuem milhares de características.

***Gradient Boosting* (GB)**

GB também é um classificador pertencente à família de AD, mas que utiliza uma técnica de aprendizado por conjunto [Alfaiz and Fati, 2022]. Diferentemente de FA, que utiliza o conceito de empacotamento (do inglês, *bagging*) e as árvores são construídas de forma independente uma das outras, GB opera sob o conceito de aprimoramento (do inglês, *boosting*) em que o aprendizado se reforça aos poucos, permitindo o máximo de desempenho em cada iteração [Muaz et al., 2020]. O aprimoramento busca construir novas árvores a partir do erro das outras previamente treinadas, melhorando aos poucos o desempenho do algoritmo como um todo.

Outros métodos também estão inseridos no contexto do GB, como o *CatBoost* (CB), uma abreviação do termo *Categorical Boosting*, e *AdaBoost* (AB), também uma abreviação, mas para o termo *Adaptive Boosting*. No primeiro caso, também trata-se de método de aprendizado pelo qual uma série de árvores de decisão são construídas sequencialmente e cada uma delas aprende com os erros da árvore anterior à medida que o treinamento acontece [Singh et al., 2022]. A única diferença entre CB e GB está no enfoque maior que CB apresenta para otimização do tratamento de dados categóricos.

No caso do AB, é um método de aprendizado conjunto que busca melhorar o desempenho de classificadores binários por meio de evolução iterativa, como nos casos anteriores de algoritmos de aprimoramento [Baker et al., 2022]. Da mesma forma que nos casos de GB e CB, um modelo é criado a partir do

treinamento do conjunto de dados e, posteriormente, um segundo modelo é construído para melhorar o desempenho do anterior. Essa operação acontece sucessivamente até que a otimização do modelo como um todo é completada, transformando um classificador fraco (como, por exemplo, AD) em um classificador mais forte [Devarakonda et al., 2023].

***K-Nearest Neighbors* (KNN)**

KNN baseia-se em medidas de similaridade, utilizando métricas de distância e um valor K [Alfaiz and Fati, 2022]. As métricas de distância ajudam a localizar os vizinhos mais próximo de qualquer valor de entrada do conjunto de dados, enquanto as regras de distanciamento classificam a nova entrada de acordo com o rótulo binário a partir dos atributos dos vizinhos. Por fim, o valor k parametriza o número de vizinhos com os quais comparar a nova entrada a ser classificada.

Vale ressaltar que o valor K tem grande peso sobre o desempenho do algoritmo. Caso ele seja abaixo do valor ótimo, a classificação fica suscetível a um grande erro estatístico. Por outro lado, caso o valor seja acima do valor ótimo, valores muito distantes contribuem para a classificação, removendo detalhes característicos da classe a ser analisada [Tran and Dang, 2021]. Mesmo assim, é um método simples que consegue alcançar níveis elevados de desempenho para conjuntos de dados relativamente complexos, sem especificações prévias sobre a distribuição do conjunto de dados [Mrozek et al., 2020].

***Naive Bayes* (NB)**

NB é um método probabilístico que aproveita as informações de atributos do conjunto de dados para classificar instâncias de acordo com a variável-alvo [Alfaiz and Fati, 2022]. O termo *Naive* se refere ao fato de que cada atributo é tratado de forma independente da classe, com o modelo aprendendo as probabilidades condicionais de cada atributo no conjunto de treinamento, dada a sua classe [Hilal et al., 2022]. A classificação é então feita aplicando o teorema de Bayes para calcular a probabilidade de pertencimento à classe, dadas as características dos atributos.

É um classificador comumente usado para conjuntos de dados massivos devido à sua simplicidade, mas a independência entre os atributos do conjunto de dados é, para muitos domínios, possivelmente restritiva [Baker et al., 2022]. No entanto, mesmo em casos em que a suposição de independência não se verifica, obteve-se resultados satisfatórios [Hilal et al., 2022].

Regressão Logística (RL)

RL é um dos algoritmos tradicionais de aprendizado de máquina que ainda é utilizado hoje devido à sua análise rápida e ao método simples de processar as características de uma classe [Alfaiz and Fati, 2022]. É uma das abordagens mais predominantes para classificação binária, devido à sua capacidade de associar vários atributos, especialmente os de maior correlação com a variável-alvo, e sua habilidade de

se ajustar a diferentes atributos dependentes dessa variável-alvo [Tran and Dang, 2021; Alfaiz and Fati, 2022]. A RL estima a probabilidade de uma transação ser fraude e a decisão final é feita comparando esse valor com um limiar de corte (por exemplo, 0,5) [Tran and Dang, 2021].

II.6 Métricas de Desempenho

O ponto de partida para análise de métricas de desempenho, no caso de problemas de classificação binária - como no contexto de fraude -, é a matriz de confusão [Tran and Dang, 2021]. Nela, uma tabela é construída para descrever a diferença entre os valores verdadeiros do conjunto de dados e os valores previstos pelo modelo sobre esse mesmo conjunto [Hilal et al., 2022]. A Tabela II.1 ilustra isso: Verdadeiros Positivos (VP) são transações previstas como fraudulentas e que de fato o são, ao passo que Verdadeiros Negativos (VN) representam o caso em que são previstas como genuínas e de fato o são. Falsos Positivos (FP) e Falsos Negativos (FN) mostram as divergências entre previsão e realidade. Os FP representam o número de transações genuínas classificadas como fraudulentas, ao passo que os FN são transações fraudulentas, mas que foram consideradas genuínas [Hordri et al., 2018; Sorournejad et al., 2016].

De forma geral, para cada uma dessas equações: (i) a acurácia representa a precisão geral do modelo em classificar transações como fraudulentas e legítimas (Equação II.1); (ii) *recall*, a capacidade do modelo de identificar todos os casos positivos - isto é, transações fraudulentas -, minimizando os FN (Equação II.2); (iii) a precisão, a confiabilidade das transações identificadas como fraudulentas pelo modelo (Equação II.3); (iv) a especificidade, a capacidade do modelo em identificar corretamente as transações legítimas como não fraudulentas (Equação II.4); (v) a taxa de FP, o número de transações incorretamente classificadas como fraudulentas (FP) em relação ao total de transações que são, de fato, legítimas (Equação II.5)e (vi) escore F_1, a combinação entre precisão e *recall*, balanceando a capacidade do modelo em classificar acertadamente as transações legítimas com a sua capacidade em identificar as transações fraudulentas (Equação IV.1).

	Previsto Positivo	**Previsto Negativo**
Real Positivo	VP	FN
Real Negativo	FP	VN

Tabela II.1: Matriz de confusão

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{II.1}$$

$$Recall\ (ou\ Sensibilidade,\ ou\ \text{Taxa de VP}) = \frac{VP}{VP + FN} \tag{II.2}$$

$$Precisão = \frac{VP}{VP + FP} \tag{II.3}$$

$$Especificidade\ (ou\ \text{Taxa de VN}) = \frac{VN}{VN + FP} \tag{II.4}$$

$$\text{Taxa de FP} = \frac{FP}{FP + VN} \tag{II.5}$$

$$escore\ F_1 = \frac{2 * VP}{2 * VP + FP + FN} \tag{II.6}$$

Também é possível monitorar a variação de desempenho do modelo em termos do *trade-off* entre *recall* e precisão, o qual se caracteriza pelo simples fato de que, quando a precisão aumenta, o *recall* tende a diminuir e vice-versa [Alfaiz and Fati, 2022]. Esse monitoramento ocorre por meio da construção das curvas *Receiver Operating Characteristic* (ROC) e precisão-*recall* e, consequentemente, do cálculo de *Area Under the ROC Curve* (AUC) e de *Area Under the Precision-Recall Curve* (AUPRC) [Hilal et al., 2022]. Na curva ROC, o eixo x é representado pela taxa de FP e o eixo y, pela taxa de VP; já no caso da curva de precisão-*recall*, o eixo x é representado por *recall* e o eixo y, pela precisão.

Finalmente, no contexto de detecção de fraudes e seus conjuntos de dados desbalanceados, as métricas de desempenho de algoritmos de aprendizado de máquina acima precisam ser implementadas com cautela. Nesse tipo de situação, por exemplo, a acurácia pode levar a conclusões precipitadas sobre o algoritmo, já que uma classificação generalizada de transações como genuínas pode levar a um alto desempenho [Bhattacharyya et al., 2011]. Nesse caso, um algoritmo mal avaliado e com posterior implementação pode causar danos às empresas envolvidas, já que previsões errôneas sobre transações fraudulentas (falsos positivos) podem gerar impacto financeiro [Correa Bahnsen et al., 2016]. Por outro lado, falsos negativos também geram custos às atividades do departamento anti-fraude, por demandarem análise manual da equipe envolvida [Bhattacharyya et al., 2011]. Assim, métricas que consideram o desempenho para além da acurácia, costumam ser mais recomendadas.

II.7 Validação Cruzada

Em muitos casos, uma única iteração para avaliação de um modelo de aprendizado de máquina pode não fornecer valores precisos para as métricas de desempenho empregadas [Alfaiz and Fati, 2022]. Assim, um método que aborde múltiplas iterações para garantir uma avaliação mais confiável é recomendado.

Nesse caso, a validação cruzada revela-se uma técnica estatística útil [Amit Singh and Tiwari, 2022]. No caso da validação cruzada *k-fold*, o conjunto de dados é dividido em k partes, que representam a

fração $\frac{1}{k}$ do total do conjunto de dados. Além disso, há a utilização da validação com estratificação, em que o termo "estratificação" indica que cada *fold* apresenta uma razão igual entre transações fraudulentas e genuínas como visto no conjunto de dados original [Kaur et al., 2022]. A Figura II.2 apresenta um modelo de validação cruzada com estratificação.

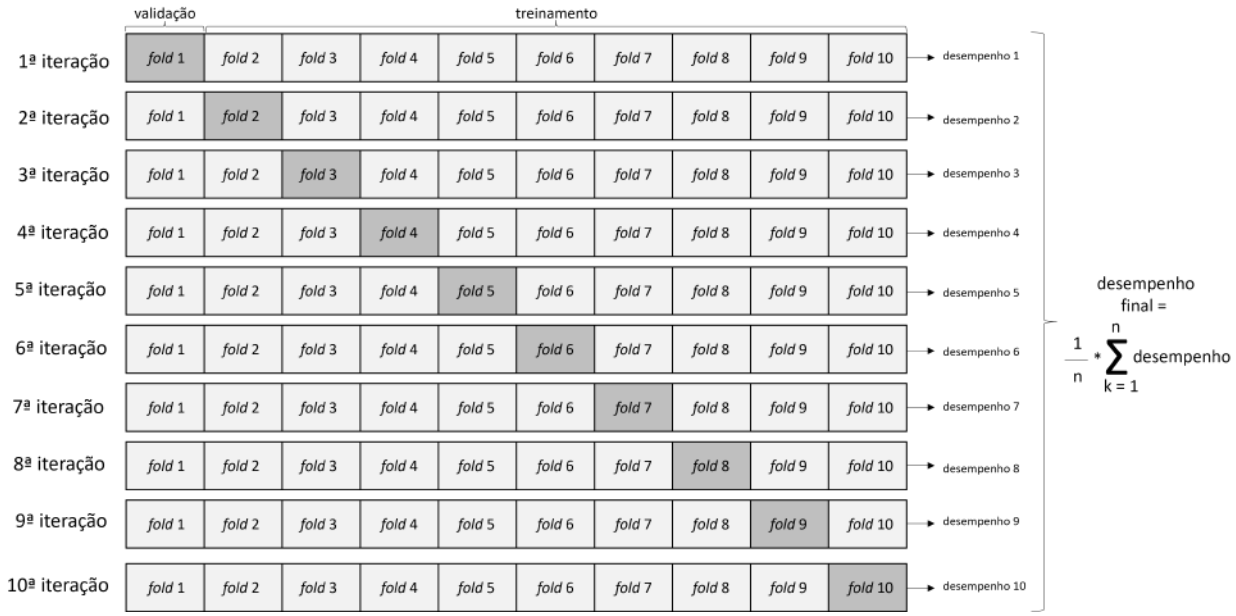


Figura II.2: Validação cruzada estratificada, com dez iterações (k = 10). Os *folds* mais escuros representam os conjuntos de validação e os mais claros, os de treinamento. O desempenho final do treinamento é calculado pela média de desempenho, em cada iteração (adaptado de [Alfaiz and Fati, 2022] e [Liu et al., 2021])

II.8 Busca por Limiar Ótimo

Em problemas de classificação binária, a saída do modelo geralmente é uma probabilidade ou um escore de confiança [Leevy et al., 2023; Johnson and Khoshgoftaar, 2019]. O limiar de decisão é o valor que determina a qual classe uma entrada será atribuída - no caso de detecção de fraudes, trata-se de uma transação fraudulenta ou genuína.

Nesse sentido, a busca por limiar ótimo refere-se ao processo de encontrar o valor de limiar que otimiza o desempenho do modelo, o qual pode ser medido por meio de métricas como precisão ou escore F_1 [Prabha and Priscilla, 2024; Leevy et al., 2023; Johnson and Khoshgoftaar, 2019]. Por padrão, um limiar comum é 0,5: isso quer dizer que, se a probabilidade ou escore de confiança for maior que 0,5, a instância é atribuída à classe positiva (transação fraudulenta); caso contrário, é atribuída à classe negativa (transação genuína). No entanto, em alguns casos, pode ser benéfico ajustar esse limiar para melhorar o desempenho do modelo conforme os requisitos específicos do problema (por exemplo, em cenários onde o custo de um falso negativo é mais severo que o de um falso positivo).

Assim, a busca pelo limiar ótimo envolve experimentar diferentes valores de limiar e avaliar o desem-

penho do modelo em cada configuração. Isso pode ser feito usando validação cruzada ou dividindo os dados em conjuntos de treinamento e teste [Prabha and Priscilla, 2024; Leevy et al., 2023; Johnson and Khoshgoftaar, 2019]. O objetivo é encontrar o limiar que equilibre as taxas de falsos positivos e falsos negativos, ou otimize a métrica desejada para o problema específico, também funcionando como um fator de ajuste em conjuntos de dados desbalanceados, além das próprias técnicas de balanceamento de classes [Leevy et al., 2023].

Capítulo III Trabalhos Relacionados

Este capítulo apresenta uma análise, com base em uma busca sistemática (Seção III.1) de trabalhos relevantes, envolvendo os temas de desbalanceamento de dados e detecção de fraudes de cartão de crédito. O objetivo é compreender o estado da arte desses temas, comparando o presente trabalho de pesquisa com outros no contexto onde se insere e identificando lacunas de conhecimento que possam ser preenchidas.

Essas lacunas podem residir em tópicos como: (i) utilização de múltiplos conjuntos de dados para validação da proposta (Seção III.2); (ii) utilização de técnicas de seleção de atributos (Seção III.3); (iii) tipos de abordagem para implementação de técnicas de balanceamento (Seção III.4); (iv) técnicas de estratificação de conjuntos de treino e teste e validação durante o treinamento (Seção III.5); (v) implementação de múltiplos classificadores para validação da proposta (Seção III.6); (vi) métricas de desempenho apropriadas para desbalanceamento (Seção III.7) e (vii) síntese dos trabalhos (Seção III.8), explicitando as lacunas em que o presente trabalho pode ou não contribuir.

III.1 Busca Sistemática

O mapeamento sistemático foi feito na base SCOPUS, que se trata de um extenso e multidisciplinar repositório científico, e considera, na sua busca, títulos, palavras-chave e resumos de trabalhos existentes na base [Carrera-Rivera et al., 2022].

Foram realizadas duas buscas no contexto do mapeamento, ambas feitas em março de 2024, mas que diferem em relação aos termos de busca utilizados. A primeira utilizou a sequência `TITLE-ABS-KEY(`"*credit card fraud detection*" `AND` "*imbalanced data*"), obteve 86 resultados e teve 13 trabalhos selecionados, ao passo que a segunda utilizou a sequência `TITLE-ABS-KEY(`"*credit card fraud detection*" `AND` "*class imbalance*"), obteve 103 resultados e teve 17 trabalhos selecionados. Vale ressaltar que, dos 103 resultados, 4 já tinham sido englobados nos 13 anteriores e, portanto, foram descartados. A Figura III.1 ilustra o processo empregado para essas buscas. De posse desse conjunto de trabalhos, iniciou-se sua análise.

Sobre o mapeamento, ele demonstra concentração de publicações e referências nos anos mais recentes, especialmente no período de 2017 a 2024. Muitos dos artigos mais recentes foram publicados em 2023 e até janeiro/março de 2024, confirmando o interesse crescente e a atualidade do campo. Com relação à inclusão e à exclusão de artigos na amostra final, o principal critério foi, a partir da leitura do

resumo de cada um deles, sua relação com o tema de detecção de fraudes em compras *online* com cartão de crédito.

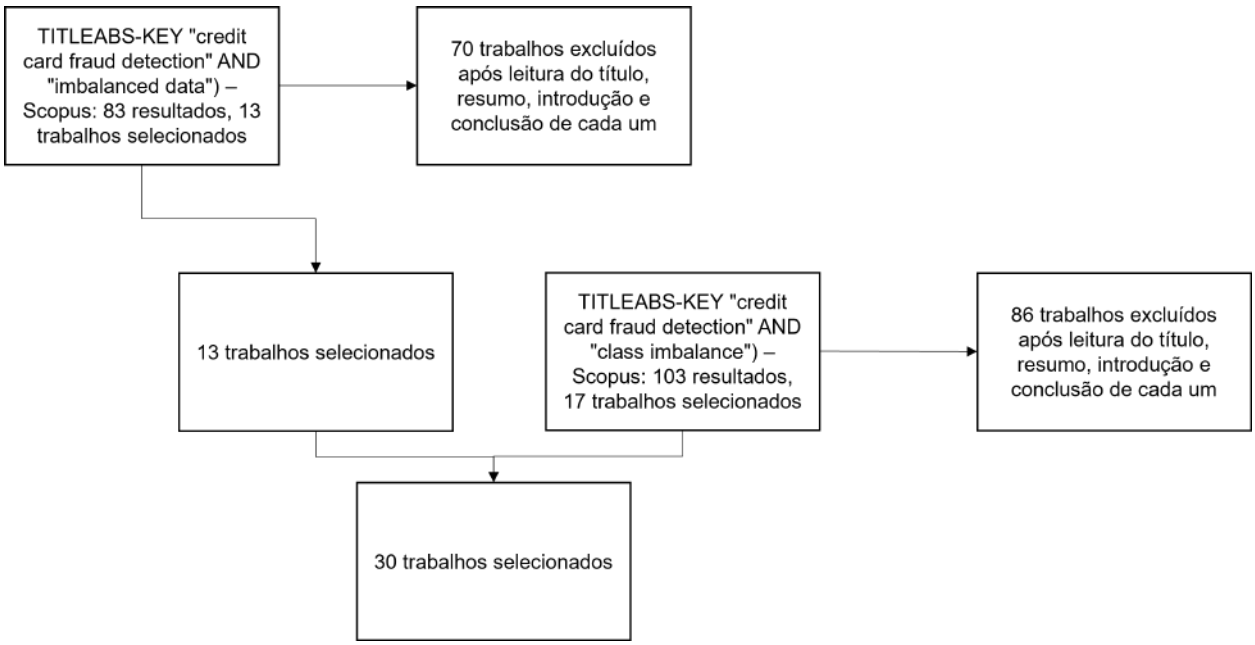


Figura III.1: Mapeamento sistemático utilizado para a busca de trabalhos relevantes. Sobre a exclusão de trabalhos, o principal motivo foi a pouca proximidade com os temas desbalanceamento de classes e detecção de fraudes em cartões de crédito.

III.2 Organização do Conjunto de Dados

A utilização de mais de um conjunto de dados para a validação de soluções em detecção de fraudes em cartão de crédito não é uma prática comum nos artigos analisados. A maioria dos estudos concentra-se na aplicação e avaliação de seus métodos em um único conjunto de dados [Strelcenia and Prakoonwit, 2023; Mahajan et al., 2023; Parekh et al., 2021].

Adicionalmente, alguns autores optam pela validação de suas soluções em múltiplos conjuntos de dados, um recurso que confere maior generalização às técnicas de detecção de fraudes em cartões de crédito [Wang, 2019; Zhang and Huang, 2020; B et al., 2022].No entanto, apesar de se mencionar a utilização de múltiplos conjuntos de dados para validar seus resultados, alguns casos não fornecem detalhes específicos sobre esses conjuntos [Strelcenia and Prakoonwit, 2023; Zhang and Huang, 2020; Xie et al., 2021].A falta de informações detalhadas sobre esses múltiplos conjuntos de dados dificulta a análise comparativa e a replicação dos experimentos.

Consequentemente, outro aspecto importante é se os conjuntos de dados estão disponíveis publicamente. Dados públicos facilitam a reprodutibilidade dos experimentos, como os disponibilizados em [Mahajan et al., 2023; Parekh et al., 2021; Cheah et al., 2023].No entanto, entre os estudos que utilizam múltiplos conjuntos de dados, observa-se variação na disponibilização pública dos dados e no nível de detalhamento dos experimentos. Wang [2019]; Shen et al. [2018]; Sapra et al. [2023].não apre-

sentam dados disponíveis publicamente, enquanto o trabalho de [Zhang and Huang, 2020] possui dados públicos apenas para o conjunto de validação e os estudos de [B et al., 2022; Xie et al., 2021; Ileberi et al., 2021] possuem dados públicos apenas para o conjunto principal de treinamento.

III.3 Utilização de Técnicas de Seleção de Atributos

Como ocorre na utilização de mais de um conjunto de dados, a utilização de técnicas de seleção de atributos também não é uma prática comum nos artigos analisados. Alguns artigos destacam e utilizam essas técnicas de seleção, mas a maioria não aplica esses métodos

Li et al. [2022] mencionam o uso de uma abordagem baseada em *Extreme Gradient Boosting* (XGBoost) para selecionar atributos, enquanto Benchaji et al. [2018] declaram que *Autoencoder* tem sido considerado uma das melhores técnicas de seleção de atributos, sendo aplicado para caracterizar mais precisamente instâncias fraudulentas. Por fim, B et al. [2022] utiliza a matriz de correlação para identificar e descartar atributos que não demonstram relação com a variável dependente, enquanto Prabha and Priscilla [2024] e Xie et al. [2021] declaram utilizar técnicas de seleção de atributos, mas sem deixar claro quais elas seriam.

III.4 Abordagem para as Técnicas de Balanceamento de Classes

Os trabalhos relacionados podem ser divididos em diferentes abordagens para a implementação das técnicas de balanceamento de classes, sendo eles: (i) proposta de novas técnicas de balanceamento, comparando seu desempenho com técnicas já conhecidas (como SMOTE e RUS), além da proposta de novas técnicas de balanceamento, mas somente comparando-as entre si (Seção III.4.1); (ii) proposta de novas técnicas de balanceamento, mas sem comparativo (Seção III.4.2); (iii) comparativo entre técnicas conhecidas, mas sem abordagem inovadora (Seção III.4.3) e (iv) utilização de uma técnica conhecida, mas sem comparativo (Seção III.4.4).

III.4.1 Proposta de Novas Técnicas de Balanceamento com Comparativo

Nesse caso, um dos principais desafios enfrentados é assegurar que as melhorias obtidas apresentem desempenho melhor do que técnicas já consagradas. Por isso, as contribuições devem passar pela validação das inovações, por meio da identificação de cenários em que técnicas tradicionais podem falhar. Há também, nesse caso, o fato de comparativo entre as técnicas inovadoras propostas por cada um dos trabalhos. Embora o comparativo entre técnicas inovadoras seja fundamental, ele impõe desafios.

Assim, no primeiro caso, Fajardo et al. [2021] exploram o uso de *Generative Adversarial Network* (GAN) em conjunto com técnicas de *oversampling*, como SMOTE, para gerar dados sintéticos que complementam as amostras da classe minoritária, garantindo maior diversidade e representatividade. Com-

parativamente, destaca-se por superar métodos tradicionais como SMOTE isolado, segundo os autores, ao evitar problemas de sobreajuste e melhorar a qualidade dos dados minoritários. Embora inovadora, a abordagem enfrenta desafios como o custo computacional elevado e o risco de "modo colapso", onde GAN podem gerar amostras redundantes.

Zhang et al. [2019] também apresentam um método inovador, chamado *Gaussian Mixture Undersampling* (GMUS), que utiliza modelos de mistura gaussiana para identificar instâncias próximas às bordas das classes, maximizando a relevância das amostras selecionadas. Em comparação com RUS, GMUS supera os desafios de eliminar dados potencialmente importantes, fornecendo uma abordagem mais precisa e informativa. Esse método se destaca pela precisão em cenários de alto desbalanceamento, mas requer ajustes cuidadosos no modelo Gaussiano para evitar sub-representação de dados importantes.

Finalmente, Benchaji et al. [2018] propõem a integração de um modelo de aprendizado profundo com a técnica *Navo Minority Oversampling Technique* (NSMOTE) para lidar com o desbalanceamento em conjuntos de dados de fraudes de cartões de crédito. NSMOTE se destaca por gerar instâncias sintéticas da classe minoritária com base em padrões mais representativos identificados por algoritmos de aprendizado profundo. Comparado a técnicas tradicionais como SMOTE, NSMOTE se diferencia pela integração direta com redes neurais, possibilitando uma representação mais precisa e adaptada ao modelo. No entanto, apresenta limitações como alta demanda computacional devido à integração com redes neurais profundas, além do risco de sobreajuste em dados minoritários escassos ou ruidosos.

No segundo caso, o desafio do custo computacional elevado das comparações ocorre pelo uso de técnicas avançadas e a necessidade de múltiplos experimentos. A principal contribuição reside em oferecer um panorama mais claro das vantagens relativas entre as abordagens propostas, ajudando na escolha de estratégias para aplicações práticas.

Dessa forma, Sethia et al. [2018] exploram a aplicação de variações inovadoras de GAN para a geração de dados sintéticos, tais como *Vanilla GAN*, *Least Squares GAN*, *Wasserstein GAN* e *Relaxed Wasserstein GAN*. As arquiteturas propostas mostram uma melhoria significativa no desempenho de classificação, com *Relaxed Wasserstein GAN* apresentando melhor desempenho que as outras. No entanto, essas técnicas enfrentam as já mencionadas limitações relacionadas aos custos computacionais elevados devido à necessidade de treinamento prolongado para alcançar estabilidade. O principal desafio, então, é equilibrar a complexidade do treinamento com os ganhos de desempenho, especialmente em aplicações em tempo real.

Cheah et al. [2023] investigam métodos inovadores híbridos, que combinam SMOTE e GAN para lidar com o desbalanceamento de classes. A técnica proposta, *GANified-SMOTE*, utiliza amostras geradas por GAN como base para a interpolação realizada pelo SMOTE, superando limitações de ambas as técnicas. Além disso, são testados outros híbridos, como *SMOTified-GAN* e *SMOTE+GAN*, mostrando

que *GANified-SMOTE* oferece o melhor desempenho. Apesar das vantagens, modelos híbridos podem dificultar a replicabilidade do *pipeline*, especialmente se não houver compartilhamento de código e dados. Assim, é fundamental abordar essa limitação ao propor um estudo sistemático e abrangente de combinações de técnicas que visa explicitamente criar um *pipeline* que permita a reprodutibilidade e ampliação das avaliações.

III.4.2 Proposta de Novas Técnicas de Balanceamento sem Comparativo

Abordagens que introduzem novas técnicas sem comparações diretas enfrentam o desafio de validar a relevância de suas propostas no contexto de outras soluções. Embora possam apresentar contribuições significativas, a ausência de um referencial comparativo pode limitar a aplicabilidade e a aceitação dessas abordagens.

Nesse sentido, Mienye and Sun [2023] utilizam o híbrido de SMOTE e *Edited Nearest Neighbours* (ENN) para balancear os dados, chamado de *Synthetic Minority Oversampling Technique-Edited Nearest Neighbours* (SMOTE-ENN), o qual oferece a geração de amostras sintéticas e a remoção de ruídos. A abordagem atinge desempenho relevante em experimentos, mas enfrenta desafios relacionados à complexidade computacional e ao risco de sobreajuste em conjuntos de dados menores.

Por sua vez, Leevy et al. [2023] adotam uma abordagem diferenciada ao focar na otimização de limiares de decisão dos classificadores, evitando alterações diretas no conjunto de dados. A técnica, aplicada em conjunto com RUS, ajusta os limiares para maximizar o desempenho, oferecendo uma alternativa eficiente e competitiva. No entanto, seu desempenho pode ser comprometido em cenários onde as distribuições de dados variam significativamente entre treino e teste. Também focada na otimização de limiares de decisão, Prabha and Priscilla [2024] apresentam essa abordagem em conjunto com algoritmos como AB, *Light Gradient-Boosting Machine* (LGBM), XGBoost, GB, FA e CB. Entre eles, o XGBoost demonstrou o melhor desempenho, alcançando resultados superiores ao limiar padrão de classificação de 0,5.

A principal contribuição em trabalhos como esses dois é a manutenção dos dados originais, melhorando significativamente o desempenho, sem introduzir redundâncias ou perda de informação associada a métodos de superamostragem ou subamostragem. O método também oferece uma solução eficiente e de baixo custo computacional, sendo uma alternativa a métodos tradicionais como SMOTE ou GAN para lidar com desbalanceamento de classes. No entanto, a técnica não aborda diretamente a melhoria da representação da classe minoritária, o que poderia ser útil em cenários com extremo desbalanceamento. Além disso, o principal desafio da técnica está na escolha e no ajuste adequado do limiar para diferentes conjuntos de dados e classificadores.

III.4.3 Comparativo entre Técnicas de Balanceamento

Apesar de não propor novas soluções, essa categoria contribui ao fornecer um entendimento aprofundado das forças e fraquezas de cada técnica conhecida em contextos variados.

Assim, Priscilla and Prabha [2020] utilizam XGBoost com ajuste de hiperparâmetros e combinado com diversas técnicas de balanceamento, como SMOTE, ADASYN, *Tomek Links* (TL), *Near Miss*, SMOTE-ENN e *Synthetic Minority Oversampling Technique-Tomek Links* (SMOTE-TL). Os melhores resultados foram observados com SMOTE-ENN, que combina a geração de dados sintéticos pelo SMOTE com a remoção de ruídos pelo ENN, resultando em uma maior precisão do modelo. SMOTE-TL também apresentou bom desempenho, eliminando dados redundantes e sobrepostos, mas técnicas como ADASYN e SMOTE apresentaram desempenho inferior devido à introdução de redundâncias em alguns cenários. *Near Miss* e TL tiveram resultados mais limitados devido à perda de dados informativos, afetando a generalização do modelo.

Muaz et al. [2020] realizam uma análise comparativa com SMOTE, *Density-Based Synthetic Minority Over-Sampling Technique* (DBSMOTE), SMOTE-ENN e RUS, com SMOTE-ENN obtendo o melhor desempenho geral, destacando-se pela combinação eficiente de geração de dados e remoção de ruídos. DBSMOTE apresentou um impacto positivo em cenários onde a diversidade da classe minoritária era essencial, mas enfrentou limitações computacionais, com RUS tendo desempenho inferior devido à eliminação aleatória de dados da classe majoritária, comprometendo a representatividade. SMOTE obteve resultados intermediários, sendo adequado para contextos onde a simplicidade e eficiência são prioritárias, mas carecendo de melhorias em cenários mais complexos.

Finalmente, Bhagwani et al. [2021] focam no uso de GAN para gerar amostras sintéticas da classe minoritária. Essa técnica demonstrou resultados superiores ao SMOTE, especialmente em termos de qualidade e diversidade das amostras geradas. Apesar disso, o uso de GAN requer alto custo computacional e uma parametrização cuidadosa para garantir estabilidade no treinamento. A técnica mostrou-se ideal para cenários onde o desbalanceamento é extremo e os recursos computacionais são adequados.

III.4.4 Aplicação de Técnicas de Balanceamento

Por fim, trabalhos que utilizam métodos consolidados sem apresentar comparações diretas enfrentam o desafio de oferecer contribuições científicas relevantes. Embora possam demonstrar a aplicação prática de técnicas conhecidas, carecem frequentemente de *insights* adicionais que avancem o estado da arte, limitando seu impacto na literatura. No entanto, podem servir como estudos exploratórios para futuros refinamentos.

Adicionalmente, Ileberi et al. [2021] utilizam o método SMOTE para lidar com o desbalanceamento de classes, combinando-o com algoritmos de aprendizado de máquina como *Support Vector Machine*

(SVM), FA, RL e XGBoost, além de integrá-los com o método AB. O uso de AB com SMOTE mostrou-se com melhor desempenho, seguido pela combinação com XGBoost. No entanto, o custo computacional dessas combinações é elevado, especialmente ao considerar o ajuste de múltiplos hiperparâmetros para evitar sobreajuste. O principal desafio reside na adaptação desse *framework* a diferentes conjuntos de dados com desbalanceamento.

No caso de Ata and Hazim [2020], adota-se uma abordagem baseada em RUS em conjunto com algoritmos NB, SVM, KNN e FA, com a própria FA apresentando o melhor desempenho geral, seguido por KNN. NB teve resultados menos expressivos devido à sua suposição de independência entre os atributos, o que não é comum em conjuntos de dados complexos. Apesar de eficiente, RUS enfrenta desafios ao remover dados da classe majoritária, o que pode levar à perda de informações relevantes, impactando a generalização dos modelos para novos conjuntos de dados.

III.5 Técnicas de Estratificação de Conjuntos de Dados e Validação Cruzada

Karthika and Senthilselvi [2023] e Ghaleb et al. [2023] indicam que a detecção de fraudes em transações online é um desafio, sobretudo devido ao desbalanceamento de classes, onde o número de transações fraudulentas é bem menor que o número de transações genuínas [Ileberi et al., 2021]. Decorrente desse desbalanceamento, Leevy et al. [2023] citam a importância da validação cruzada com estratificação durante a classificação para melhorar a capacidade de generalização, assim como Muaz et al. [2020] citam o mesmo benefício para a estratificação durante a separação dos conjuntos de treino e teste.

Dessa forma, diferentes abordagens para técnicas de estratificação de conjuntos de dados de treino e teste e validação cruzada durante o treinamento são analisadas nos artigos selecionados. Em sua grande maioria, os trabalhos não realizam nem validação nem estratificação, utilizando particionamentos simples de treino e teste [Strelcenia and Prakoonwit, 2023; Wang, 2019; Zhang and Huang, 2020].No geral, nesses trabalhos, as técnicas de balanceamento de classes são implementadas sem que a representatividade dos conjuntos de treino e teste tenha sido explicitamente avaliada, o que pode levar a conjuntos de treino e teste com distribuições desiguais.

Um segundo grupo realiza apenas validação durante o processo de treinamento dos modelos, como o uso de validação cruzada *k-fold* para avaliar seu desempenho, mas sem preocupação com a estratificação [Mahajan et al., 2023; Mienye and Sun, 2023; Xie et al., 2021].A validação cruzada melhora a consistência dos resultados e permite uma avaliação mais confiável dos modelos. Contudo, a ausência de estratificação no particionamento inicial pode ter introduzido variabilidade nas distribuições de treino e teste, prejudicando a representatividade dos dados e potencialmente levando a resultados menos generalizáveis.

Há ainda estudos que realizam apenas estratificação, garantindo a representatividade proporcional das classes no particionamento de treino e teste, mas sem aplicar validação durante o treinamento [Pa-

rekh et al., 2021]. A estratificação garante que ambas as classes (majoritária e minoritária) sejam representadas proporcionalmente em cada conjunto de treinamento e teste, ajudando a evitar viés nas distribuições. No entanto, a ausência de validação cruzada limita o desempenho dos resultados.

Por fim, alguns combinam ambas as abordagens, aplicando estratificação e validação cruzada, criando *pipelines* de melhor desempenho, embora com custo computacional mais elevado [Berkmans and Karthick, 2022; Muaz et al., 2020]. Essa combinação permite maximizar a confiabilidade dos resultados, reduzindo o impacto de variações nos conjuntos de dados e o desbalanceamento extremo de classes. Contudo, essa abordagem apresenta alta complexidade computacional, limitando sua escalabilidade para conjuntos de dados muito grandes.

III.6 Implementação de Múltiplos Classificadores

Ileberi et al. [2021], Priscilla and Prabha [2020], Zhang et al. [2019], Muaz et al. [2020] e Xie et al. [2021] mostram a importância de utilizar diversos classificadores em modelos de detecção de fraudes para avaliação de desempenho. Assim, os trabalhos analisados podem ser classificados em três diferentes categorias: (i) implementação de mais de um classificador para avaliação das técnicas de balanceamento avaliadas; (ii) implementação de somente um classificador; e (iii) implementação de um classificador não especificado.

Strelcenia and Prakoonwit [2023]; Wang [2019]; Zhang and Huang [2020]se encontram no primeiro grupo, que costuma contribuir significativamente para a avaliação detalhada da eficácia de técnicas de balanceamento. No entanto, enfrentam limitações devido ao custo computacional elevado e ao desafio de ajustar múltiplos hiperparâmetros, especialmente em modelos mais complexos. B et al. [2022], por exemplo, ressaltam que o uso de classificadores variados pode aumentar o desempenho da avaliação, porém enfrentam desafios de custo computacional.

Trabalhos que empregam apenas um classificador apresentam contribuições na simplicidade metodológica e facilidade de replicação, mas enfrentam limitações na avaliação devido à ausência de comparativos com outros classificadores Mahajan et al. [2023]; Fajardo et al. [2021]; Bhagwani et al. [2021].A exploração de múltiplos classificadores poderia enriquecer as análises e aumentar a aplicabilidade dos resultados em diferentes contextos. Em Mahajan et al. [2023] e Sethia et al. [2018], por exemplo, a ausência de comparação com outros classificadores limitou a abrangência das conclusões.

Por outro lado, estudos que não especificam os classificadores utilizados representam um desafio significativo, comprometendo a reprodutibilidade e dificultando a comparação com outros estudos [Shen et al., 2018; Benchaji et al., 2018; Venkateswarlu et al., 2022] . Sem informações sobre os classificadores, não é possível avaliar se os ganhos observados decorrem do algoritmo de balanceamento ou de características específicas do modelo.

III.7 Métricas de Desempenho apropriadas para Desbalanceamento

Em problemas com conjuntos de dados desbalanceados, a escolha de métricas adequadas é essencial para uma avaliação justa do desempenho do modelo. Métricas como escore F_1 e *recall* são apropriadas porque capturam o desempenho equilibrado entre classes ou priorizam a classe minoritária, que geralmente é o foco de interesse. Por outro lado, métricas como acurácia e precisão, isoladamente, oferecem limitações, pois podem supervalorizar o desempenho na classe majoritária e mascarar o real impacto na classe minoritária. Além disso, ferramentas como a matriz de confusão são fundamentais para fornecer maior entendimento sobre os erros do modelo.

A maioria dos trabalhos analisados conseguem avaliar o desempenho de seus modelos de forma apropriada, utilizando o escore F_1 [Zhang et al., 2019], ou o *recall* [Wang, 2019; Shen et al., 2018; Baabdullah et al., 2020]ou ambos [Strelcenia and Prakoonwit, 2023; Parekh et al., 2021; B et al., 2022], dado o contexto do desbalanceamento de classes. Por exemplo, Strelcenia and Prakoonwit [2023] priorizam métricas como escore F_1 devido à sua relevância em conjuntos de dados desbalanceados, mas aponta a dificuldade em equilibrar *recall* e precisão. Embora amplamente utilizada, a métrica AUC pode ser limitada em cenários de alto desbalanceamento, pois pode não refletir bem o desempenho sobre a classe minoritária Zhang and Huang [2020].

Por outro lado, Mahajan et al. [2023]; Bhagwani et al. [2021] utilizam somente acurácia como métrica de desempenho. Embora a acurácia seja amplamente compreendida, sua utilização em conjuntos de dados desbalanceados é limitada, já que ela pode mascarar o impacto de falsos negativos. Benchaji et al. [2018]; Venkateswarlu et al. [2022], por sua vez, não especificam as métricas utilizadas, o que compromete a reprodutibilidade dos resultados e dificulta a análise comparativa com outros estudos.

III.8 Síntese dos Trabalhos

As Tabelas III.1 e III.2 apresentam uma síntese dos trabalhos analisados nesse capítulo, comparando-as com a proposta desse trabalho. A Tabela III.3 faz um resumo quantitativo dos 31 trabalhos analisados, incluindo o presente trabalho.

À luz dos trabalhos revisados, recomenda-se tratar seleção de atributos, balanceamento e classificação como fatores combináveis. Para os classificadores, a utilização de múltiplos classificadores como comparação para garantir a avaliação do desempenho dos modelos é válida. Nesse contexto, aqueles baseados em árvores (como AD, FA e ET) e em *boosting* (como GB e XGBoost costumam ser citados. No que tange às técnicas de balanceamento, técnicas como SMOTE, ADASYN e RUS também são citadas, com a estratégia de otimização de limiares se mostrando competitiva, já que ajusta o ponto de corte do classificador sem a necessidade de alterar diretamente o conjunto de dados. Isso evita a introdução de redundâncias ou a perda de informações.

Para seleção de atributos, manter um *baseline* com todos os atributos e, em paralelo, usar uma técnica simples e reprodutível — como QQ para variáveis categóricas e ANOVA para numéricas —, fixando um ranqueamento de melhores atributos (*e.g.*: 50 ou 150) que preserve variabilidade sem gerar ruído. Finalmente, a literatura analisada ressalta a importância metodológica da validação cruzada com estratificação e do uso de métricas como escore F_1 para avaliar de forma apropriada o desempenho do modelo.

Comparativamente aos artigos analisados, a proposta desse trabalho apresenta contribuições com relação a: (i) amplitude de avaliação, com a utilização de múltiplos conjuntos de dados, validando os resultados em diferentes cenários; (ii) metodologia ampla, ao incluir múltiplas técnicas de balanceamento de classes, estratificação dos conjuntos de treino e teste, validação cruzada durante o treinamento, implementação do limiar ótimo de classificação e a utilização de múltiplos classificadores; e (iii) pertinência de avaliação, com o uso do escore F_1, relevante em cenários de desbalanceamento extremo.

Apesar de não propor uma nova técnica de balanceamento, a principal contribuição deste trabalho reside na sua metodologia, focada na avaliação comparativa de combinações (ou *pipelines*). Cada um desses *pipelines* são abrangentes, englobando técnicas de seleção de atributos, de balanceamento de classes e a aplicação de múltiplos classificadores. A análise do desempenho dessas combinações foi validada em diferentes cenários, utilizando dois conjuntos de dados desbalanceados (melhor explicados no Capítulo IV), o que confere maior generalização aos resultados obtidos. O foco em métricas como escore F_1 garante que as diretrizes práticas fornecidas para a escolha de abordagens sejam avaliadas, especialmente sob a ótica da classe minoritária.

Tabela III.1: Comparativo dos Trabalhos Relacionados

Autor	**Múltiplos Conjuntos de Dados**	**Seleção de Atributos**	**Abordagem**	**Limiar Ótimo**	**Estratificação**	**Validação Cruzada**	**Múltiplos Classificadores**	**Métricas Apropriadas**
Alamri and Ykhlef [2024]	Não	Não	Inovadora com Comparativo	Não	Não	Não	Não	Sim (*recall* e escore F_1)
Ata and Hazim [2020]	Não	Não	Aplicação	Não	Não	Sim	Sim	Sim (*recall*)
B et al. [2022]	Sim	Sim	Comparativo	Não	Não	Não	Sim	Sim (*recall* e escore F_1)
Baabdullah et al. [2020]	Não	Não	Comparativo	Não	Não	Não	Sim	Sim (*recall*)
Benchaji et al. [2018]	Não	Sim	Inovadora com comparativo	Não	Não	Não	Não Especificado	Não Especificado
Berkmans and Karthick [2022]	Não	Não	Comparativo	Não	Sim	Sim	Sim	Sim (*recall* e escore F_1)
Bhagwani et al. [2021]	Não	Não	Comparativo	Não	Não	Não	Não	Não
Cheah et al. [2023]	Não	Não	Inovadora com comparativo	Não	Não	Não	Sim	Sim (*recall* e escore F_1)
Fajardo et al. [2021]	Sim	Não	Inovadora com comparativo	Não	Não	Não	Não	Sim (*recall* e escore F_1)
Ghaleb et al. [2023]	Não	Não	Inovadora com comparativo	Não	Não	Não	Sim	Sim (*recall* e escore F_1)
Ileberi et al. [2021]	Sim	Não	Aplicação	Não	Não	Sim	Sim	Sim (*recall* e escore F_1)
Karthika and Senthilselvi [2023]	Sim	Não	Inovadora com comparativo	Não	Não	Não	Sim	Sim (*recall* e escore F_1)
Leevy et al. [2023]	Não	Não	Inovadora sem comparativo	Sim	Não	Sim	Sim	Sim (*recall* e escore F_1)
Li et al. [2022]	Não	Sim	Inovadora com comparativo	Não	Não	Não	Sim	Sim (*recall*)
Mahajan et al. [2023]	Não	Não	Comparativo	Não	Não	Sim	Não	Não
Mienye and Sun [2023]	Não	Não	Inovadora sem comparativo	Não	Não	Sim	Não	Sim (*recall*)
Muaz et al. [2020]	Sim	Não	Comparativo	Não	Sim	Sim	Sim	Sim (*recall* e escore F_1)
Parekh et al. [2021]	Não	Não	Comparativo	Não	Sim	Não	Sim	Sim (*recall* e escore F_1)
Prabha and Priscilla [2024]	Não	Não	Inovadora sem comparativo	Sim	Não	Não	Sim	Sim (*recall* e escore F_1)

Tabela III.2: Comparativo dos Trabalhos Relacionados (cont.)

Autor	**Múltiplos Conjuntos de Dados**	**Seleção de Atributos**	**Abordagem**	**Limiar Ótimo**	**Estratificação**	**Validação Cruzada**	**Múltiplos Classificadores**	**Métricas Apropriadas**
Pranavi et al. [2022]	Não	Não	Aplicação	Não	Não	Sim	Sim	Sim (*recall* e escore F_1)
Priscilla and Prabha [2020]	Sim	Sim	Comparativo	Não	Não	Sim	Sim	Sim (*recall* e escore F_1)
Sapra et al. [2023]	Sim	Não	Inovadora com comparativo	Não	Não	Não	Sim	Sim (*recall* e escore F_1)
Sethia et al. [2018]	Não	Não	Inovadora com comparativo	Não	Não	Não	Não	Sim (*recall*)
Shen et al. [2018]	Sim	Não	Inovadora sem comparativo	Não	Não	Não	Não Especificado	Sim (*recall*)
Strelcenia and Prakoonwit [2023]	Não	Não	Inovadora com comparativo	Não	Não	Não	Sim	Sim (*recall* e escore F_1)
Venkateswarlu et al. [2022]	Não	Não	Inovadora com comparativo	Não	Não	Não	Não Especificado	Não Especificado
Wang [2019]	Sim	Não	Inovadora com comparativo	Não	Não	Não	Sim	Sim (*recall*)
Xie et al. [2021]	Sim	Sim	Inovadora com comparativo	Não	Não	Sim	Sim	Sim (*recall*)
Zhang and Huang [2020]	Sim	Não	Comparativo	Não	Não	Não	Sim	Sim (AUC)
Zhang et al. [2019]	Sim	Não	Inovadora com comparativo	Não	Não	Não	Sim	Sim (escore F_1)
Esta dissertação	**Sim**	**Sim**	**Comparativo**	**Sim**	**Sim**	**Sim**	**Sim**	**Sim (escore F_1)**

Tabela III.3: Resumo Quantitativo dos Trabalhos Relacionados

Critério	**Quantidade (%)**
Total de trabalhos analisados	31 (100%)
Utilizam múltiplos conjuntos de dados	13 (42%)
Utilizam seleção de atributos	6 (19%)
Tipo de abordagem	
Inovadora com comparativo	14 (45%)
Comparativo	10 (32%)
Inovadora sem comparativo	4 (13%)
Aplicação	3 (10%)
Utilizam ajuste de limiar ótimo	3 (10%)
Utilizam estratificação dos dados	4 (13%)
Utilizam validação cruzada	10 (32%)
Utilizam múltiplos classificadores	22 (71%)
Utilizam métricas apropriadas (*recall*, F_1, AUC)	27 (87%)

Capítulo IV Metodologia

Este capítulo visa expor detalhadamente a metodologia para a análise comparativa de métodos de balanceamento de classes no contexto de detecção de fraude de cartão de crédito. Ao longo deste capítulo, são discutidas as etapas da metodologia utilizada: na Seção IV.1 destacam-se os conjuntos de dados utilizados nos experimentos; nas Seções IV.2 e IV.3 são explicados os processos de ajuste do conjunto de dados e de classificação implementados; e na Seção IV.4, destaca-se o processo de avaliação de desempenho da classificação. É fundamental ressaltar que a metodologia empregada envolveu um histórico evolutivo de experimentos e combinações de técnicas. A descrição completa desse histórico, incluindo as avaliações preliminares de classificadores, métodos de seleção de atributos e as iterações que levaram aos resultados finais, é detalhada no Capítulo V, a fim de demonstrar a não-seletividade da análise e o rigor do processo de pesquisa. As etapas que compõem a metodologia estão descritos na Figura IV.1.

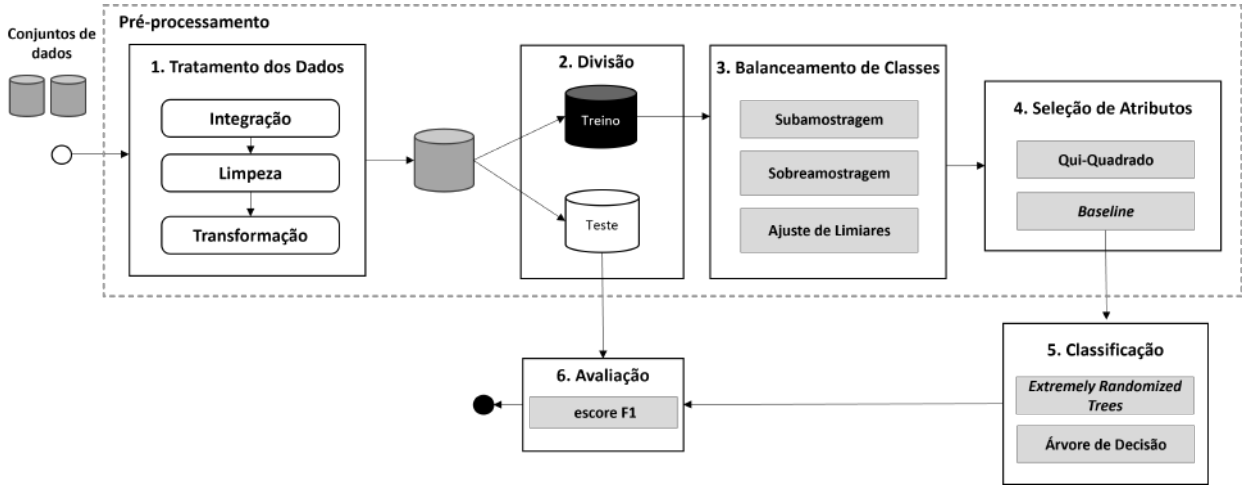


Figura IV.1: Visão geral da metodologia definida para os experimentos.

IV.1 Conjuntos de Dados

São utilizados dois conjuntos de dados públicos desbalanceados de transações de cartão de crédito. O primeiro[1] foi desenvolvido pelo *Institute of Electrical and Electronics Engineers - Computational Intelligence Society* (IEEE-CIS) em parceria com a Vesta Corporation, líder global em serviços de pagamentos. Ele apresenta cerca de 590 mil entradas (em que 3,5% das transações são fraudulentas), com 433 atribu-

[1] `https://www.kaggle.com/competitions/ieee-fraud-detection/data`

tos, e se apresenta em duas partes: (i) tabela “Transação” e (ii) tabela “Identidade”.

Mais detalhadamente, os atributos da Tabela “Transação” descritos na documentação do conjunto de dados são: (i) `TransactionID`, com o ID da transação; (ii) `TransactionDT`, sendo um dado temporal, contabilizado a partir de uma referência de data específica (não é de fato um atributo de data e horário e é necessário um tratamento posterior); (iii) `TransactionAMT`, com o valor total da transação, em dólares; (iv) `ProductCD`, sendo o código do produto de cada transação; (v) `card1 - card6`, contendo informações do cartão usado para o pagamento, como tipo (débito ou crédito) e bandeira; (vi) `addr`, contendo o endereço da fatura; (vii) `P_ and (R__) emaildomain`, com os e-mails do comprador e do recebdor da compra; (viii) `C1-C14`: valores como o número de endereços que o comprador tem cadastrado na plataforma de *e-commerce*; (ix) `D1-D15`, contendo informações de tempo, como dias entre as compras; (x) `M1-M9`, com valores como nome no cartão; e (xi) `Vxxx`, que representam atributos inseridos pela Vesta no conjunto de dados. Da Tabela “Identidade”, os atributos descritos são: (i) `TransactionID`, com o ID da transação; (ii) `DeviceType`, com o tipo de dispositivo utilizando na compra, como celular ou computador; e (iii) `DeviceInfo`, com o sistema operacional do dispositivo utilizado.

O segundo conjunto de dados[2] foi desenvolvido pelo *Worldline*[3] juntamente com o *Machine Learning Group* (MLG) da *Université Libre de Bruxelles* (ULB) e serve como conjunto de validação do *pipeline* desenvolvido com base no conjunto de dados IEEE-CIS. Ele apresenta cerca de 285 mil entradas (em que 0,2% das transações são fraudulentas) com 30 atributos, tratando-se de um conjunto de transações feito por empresas de cartão de crédito europeias em setembro de 2013, disponível com dados já tratados. Por conta de cláusulas de confidencialidade, a origem dos dados não pode ser especificada, exceto o fato de que os únicos atributos que não passaram por essa transformação são justamente os de data e valor da transação, além do próprio rótulo de transação fraudulenta ou não. Esse conjunto de dados é apenas utilizado como validação a partir de toda a metodologia empregada para o primeiro, do IEEE-CIS. Vale ressaltar também que esse conjunto passa por todo o *pipeline* já pronto, ajustado a partir do trabalho com o conjunto de dados IEEE-CIS.

IV.2 Pré-Processamento

A fase de pré-processamento desse trabalho serve, essencialmente, para a manipulação e transformação dos dados brutos em um formato adequado para a classificação. Nesse caso, isso inclui as etapas de tratamento de dados, divisão do conjunto de dados em treino e teste utilizando estratificação, balanceamento das classes (transação legítima e fraudulenta) e seleção de atributos.

[2] `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`
[3] `https://worldline.com/`

IV.2.1 Tratamento dos Dados

O tratamento de dados ocorre apenas para o conjunto de dados IEEE-CIS, dado que o conjunto MLG-ULB é apenas para validação. Dessa forma, inicia-se com a preparação dos dados, integrando as tabelas "Transação" e "Identidade" por meio do atributo `TransactionID`, obtendo um conjunto de dados unificado e completo. Em seguida, removem-se as colunas com mais de 75% dos valores faltantes - o que representa a exclusão de 208 atributos -, além da remoção das colunas com mais do que 75% dos valores únicos - o que representa a exclusão de mais 137 atributos. O limiar de 75% foi estabelecido após análise empírica, uma vez que a aplicação de limites superiores (90%, 85% e 80%) demonstrou consistentemente uma queda no desempenho dos modelos. Os 345 atributos removidos representam, em sua maioria, o conjunto incluído pela Vesta (`Vxxx`) e grande parte dos atributos da tabela "Identidade". Dessa forma, ao final, o conjunto de dados é reduzido de 433 atributos para 88, que podem ser vistos na Tabela IV.1.

Atributos Selecionados			
_Days	D4	V2	V51
_Hours	D5	V280	V52
_Weekdays	dist1	V282	V53
addr1	M1	V283	V54
C1	M2	V285	V6
C11	M3	V29	V69
C13	M4	V3	V7
C14	M5	V30	V70
C2	M6	V307	V75
C5	M7	V308	V76
C6	M8	V310	V8
C9	M9	V35	V80
card1	P_emaildomain	V36	V81
card2	ProductCD	V39	V84
card4	TransactionAmt	V4	V85
card5	V1	V40	V9
D1	V10	V42	V90
D10	V11	V43	V91
D11	V12	V48	V92
D15	V127	V49	V93
D2	V13	V5	V96
D3	V130	V50	V99

Tabela IV.1: Lista final dos atributos selecionados para treinamento (conjunto de dados utilizado: IEEE-CIS)

Por fim, é realizada a transformação dos dados, como a transformação do atributo `TransactionDT` para um formato de data e hora YYYY-MM-DD HH:MM:SS e a substituição de valores faltantes de atribu-

tos numéricos pela média e de atributos categóricos pela moda. A escolha pela moda e média é guiada pela busca de preservar a tendência central dos dados, minimizando assim o impacto sobre a distribuição global. Feita a substituição, é realizada a codificação de colunas categóricas com a técnica de codificação *one-hot*, que basicamente faz essa transformação de dados categóricos em dados numéricos [Yeturu et al., 2023]. Esse método preserva a informação da categoria original, ao mesmo tempo, em que evita introduzir ordem ou hierarquia artificial entre as categorias. Ao final dessa última etapa, o conjunto de dados fica com um total de 163 atributos.

IV.2.2 Divisão

A fase de divisão entre a base de treino (80% dos dados, tanto para o conjunto IEEE-CIS quanto para o conjunto MLG-ULB) e a base de teste (20% dos dados restantes) compreende também a estratificação da variável-alvo, aplicada exclusivamente à base de treino. Dada a disparidade entre as classes fraudulenta e não fraudulenta, a estratificação da variável-alvo é adotada como estratégia para garantir que ambas as classes estejam representadas proporcionalmente em ambos os conjuntos. Essa abordagem se revela crucial em cenários desbalanceados, onde a raridade das instâncias fraudulentas poderia levar a uma alocação insuficiente desses casos no conjunto de teste, comprometendo a avaliação do desempenho do modelo. Vale ressaltar que, após a divisão estratificada, a base de teste é imediatamente isolada e mantida inalterada. Todas as etapas subsequentes são aplicadas exclusivamente à base de treino, prevenindo o vazamento de dados que poderia ocorrer se informações da base de teste influenciassem o treinamento.

Assim, assegura-se que o modelo seja treinado e avaliado em um ambiente que reflita a distribuição da variável-alvo no conjunto de dados original. Essa consideração é particularmente relevante para evitar que o modelo seja otimizado apenas para a classe majoritária, negligenciando a capacidade de classificar adequadamente as instâncias da classe minoritária.

IV.2.3 Balanceamento de Classes

A etapa seguinte do pré-processamento consiste em aplicar técnicas para o balanceamento das classes exclusivamente sobre a base de treino estratificada. A aplicação de cada técnica (superamostragem, subamostragem ou ajuste de limiar) resulta em uma nova instância da base de treino, que é utilizada separadamente na fase de classificação. Esta metodologia garante que o modelo seja treinado em dados balanceados ou ajustados, enquanto a avaliação final é sempre realizada sobre a base de teste original.

Inicialmente, são usadas técnicas de superamostragem, como *Resample* e ROS, aumentando exemplos da classe minoritária de forma aleatória para garantir maior representatividade e diversidade dos dados, e técnicas de subamostragem como RUS, para reduzir de forma aleatória exemplos da classe majoritária. Posteriormente, técnicas de superamostragem mais avançadas, SMOTE e BSMOTE, são em-

pregadas para gerar novas amostras sintéticas, melhorando a representatividade da classe minoritária e a precisão dos modelos. No caso de técnicas avançadas de subamostragem, *Near Miss* e IHT são posteriormente implementadas, para, respectivamente, selecionar instâncias da classe majoritária que estão próximas da classe minoritária, ajudando a ajustar a fronteira de decisão entre as classes, e remover instâncias da classe majoritária que são difíceis de classificar corretamente. Finalmente, no caso da técnica avançada de superamostragem ADASYN, geram-se amostras sintéticas adaptativas que se concentram em áreas onde a classe minoritária é mais difícil de aprender.

O processo de análise de limiares de decisão também é compreendido como uma técnica de balanceamento de classes, em que ajusta diferentes limiares para a probabilidade de classificação entre transação genuína ou fraudulenta pelo modelo. Nessa etapa de balanceamento de classes, este processo é repetido para uma série de limiares definidos em um intervalo, como de 0,05 a 0,75 - escolhidos a partir de avaliação empírica de desempenho, já que a partir de 0,75, os resultados não variavam. A média das métricas de desempenho é então calculada para cada limiar, permitindo identificar o limiar que proporciona a melhor métrica de desempenho.

Vale ressaltar que as técnicas de superamostragem chegaram, ao final da sua implementação, a uma proporção de 50/50, com 570 mil entradas fraudulentas e 570 mil entradas genuínas no conjunto de dados IEEE-CIS e 240 mil para cada caso no conjunto de dados MLG-ULB. Por outro lado, as técnicas de subamostragem chegaram a uma proporção de 90/10, com 230 mil entradas genuínas e 20 mil entradas fraudulentas no conjunto IEEE-CIS e 4,9 mil entradas genuínas e 492 entradas fraudulentas no conjunto MLG-ULB. O ajuste de limiares não gerou balanceamento, ficando com a proporção de 590 mil transações genuínas e 20 mil fraudulentas no conjunto IEEE-CIS e 285 mil transações genuínas e 492 transações fraudulentas no conjunto MLG-ULB.

IV.2.4 Seleção de Atributos

Por fim, a última etapa do pré-processamento realiza a seleção de atributos mais relevantes para a detecção de fraudes, tanto para os conjuntos de dados superamostrados quanto para os subamostrados. Para isso, utiliza-se a técnica QQ, em que se avalia a independência entre cada atributo e a variável alvo, permitindo a identificação de atributos que possuem maior correlação com a ocorrência de fraudes [Laborda and Ryoo, 2021]. A técnica é aplicada sobre cada um dos métodos de balanceamento de classes previamente descritos. Para o conjunto de dados IEEE-CIS, foram selecionados os 100 atributos de maior valor pela técnica mencionada e, no caso, do conjunto MLG-ULB, os primeiros 20 atributos - empiricamente, foram os patamares de 100 e 20 que geraram melhores desempenhos do modelo. Para fins de comparação, também é considerada na metodologia a abordagem sem seleção de atributos (*baseline*). Vale ressaltar que todas as combinações de técnicas de balanceamento, seleção de atributos e classificadores foram inicialmente testadas e avaliadas no conjunto de dados IEEE-CIS, que serviu como base

para a formulação das diretrizes de desempenho. O conjunto MLG-ULB foi reservado exclusivamente para a etapa final de validação, a fim de verificar a generalização das diretrizes estabelecidas no IEEE-CIS.

IV.3 Treinamento do Classificador

Essa etapa utiliza os classificadores ET e AD pois, ao longo dos experimentos iniciais (a serem detalhados no Capítulo V), foram os que demonstraram melhor desempenho nos conjuntos de treino e teste da base IEEE-CIS, em relação a classificadores como FA, CB, GB, AB, RL, KNN e NB. Assim, ambos são implementados, inicialmente, no caso de conjuntos de dados sem balanceamento de classes, aplicando tanto a técnica de seleção QQ quanto o *baseline*. Posteriormente, esses classificadores são combinados com as técnicas de balanceamento e técnicas de seleção de atributos citadas. Todas as combinações foram testadas no conjunto IEEE-CIS para identificar as melhores configurações (*pipelines*). Somente os *pipelines* que demonstraram o melhor escore F_1 médio no IEEE-CIS foram replicados no conjunto MLG-ULB para validar se os resultados e as diretrizes de desempenho convergiam em uma base de dados distinta. Vale citar que classificação é realizada utilizando validação cruzada com estratificação do conjunto de dados, sendo esta aplicada estritamente sobre a base de treino.

IV.4 Avaliação

A avaliação dos classificadores é realizada pela métrica escore F_1 (Equação IV.1), considerando tanto as fraudes não detectadas quanto as transações legítimas erroneamente classificadas como fraude. Como existe estratificação do conjunto de dados, calcula-se uma média entre os escores F_1 de cada *fold*, o que denomina a métrica final como escore F_1 médio.

$$F_1 = \frac{2 * VP}{2 * VP + FP + FN} \tag{IV.1}$$

em que VP, FP e FN são citados em II.6.

Capítulo V Resultados

Nesta seção, apresentamos os resultados obtidos a partir dos experimentos executados segundo a metodologia descrita no capítulo anterior. A Seção V.1 apresenta configurações de *hardware* e *software* utilizadas para execução dos experimentos, sendo seguida pela Seção V.2, que analisa o conjunto de dados principal desse trabalho, e Seção V.3, que mostra a linha do tempo dos experimentos executados. A Seção V.4 apresenta uma síntese dos melhores resultados alcançados.

V.1 Configuração dos Experimentos

Os experimentos foram realizados em um computador com processador Intel(R) Core(TM) i5-10400 CPU de 2,90GHz e 128 GB de memória RAM sob o sistema Ubuntu v20.04, e implementados[1] em Python usando as bibliotecas *pandas*[2], *scikit-learn*[3], *numpy*[4], *imbalanced-learn*[5], *lightgbm*[6], *xgboost*[7] e *catboost*[8]. Em ambos os classificadores, ET e AD, foi utilizada a validação cruzada *k-fold* com 10 *folds*. No classificador ET, em específico, foram utilizadas 100 árvores como estimadores. Os demais parâmetros dos classificadores e técnicas de balanceamento foram os valores padrões das funções nas respectivas bibliotecas.

V.2 Análise Exploratória da Base IEEE-CIS

A *Análise Exploratória de Dados* (AED) visa identificar padrões e características relevantes que possam contribuir para a construção de modelos preditivos mais eficientes [Alamri and Ykhlef, 2024]. Nesta seção, realiza-se uma exploração detalhada de várias variáveis presentes na base de dados IEEE-CIS, após à limpeza da base, focando em aspectos como o tipo de dispositivo utilizado, valor em dinheiro das transações, características temporais e geográficas, além de outros fatores que podem influenciar a ocorrência de fraudes.

Um dos primeiros aspectos analisados é o impacto do tipo de dispositivo utilizado nas transações,

[1] O código-fonte está disponível em `https://github.com/arthur-silveirasc/credit_card_fraud_detection`.
[2] `https://pandas.pydata.org/pandas-docs/version/2.0.2/index.html`
[3] `https://scikit-learn.org/stable/whats_new/v1.2.html`
[4] `https://numpy.org/doc/stable/release/1.24.0-notes.html`
[5] `https://imbalanced-learn.org/stable/`
[6] `https://lightgbm.readthedocs.io/en/latest/Python-Intro.html`
[7] `https://xgboost.readthedocs.io/en/stable/python/python_intro.html`
[8] `https://catboost.ai/en/docs/concepts/installationl`

em que as fraudes estão mais associadas a transações feitas via dispositivos móveis, sugerindo que usuários de dispositivos móveis podem estar mais suscetíveis a fraudes ou que fraudadores preferem realizar transações fraudulentas por meio desses dispositivos. Nota-se também que dispositivos com sistemas operacionais específicos, como Windows e iOS, são os mais utilizados para transações. Entretanto, essa análise por si só não indica uma correlação direta com a fraude, mas destaca a predominância de certos dispositivos na base de dados.

Quanto aos valores das transações (em dólares), observa-se que a maioria se concentra em montantes baixos, com mediana de aproximadamente 69 dólares. Quando os dados são segmentados entre transações fraudulentas e não fraudulentas, percebe-se que as transações fraudulentas tendem a envolver montantes ligeiramente mais altos em comparação às transações genuínas, com uma média de 149 dólares contra 134 dólares, respectivamente.

A análise da distribuição do atributo `ProductCD` - variável categórica que representa o tipo de produto envolvido na transação - mostra que, além do produto "C" apresentar maior número de transações fraudulentas em proporção, como mostrado na Figura V.1, esse produto também apresenta a maior proporção de fraudes, em termos de valores em dólares com cerca de 13% do valor total alocado nesse tipo de transação. Apesar da análise de *boxplots* da Figura V.2 mostrar que as médias das transações (linhas em branco) envolvendo os outros tipos de produtos são mais elevadas, o produto "C" apresenta maior risco devido ao seu alto percentual de transações fraudulentas, comparativamente.

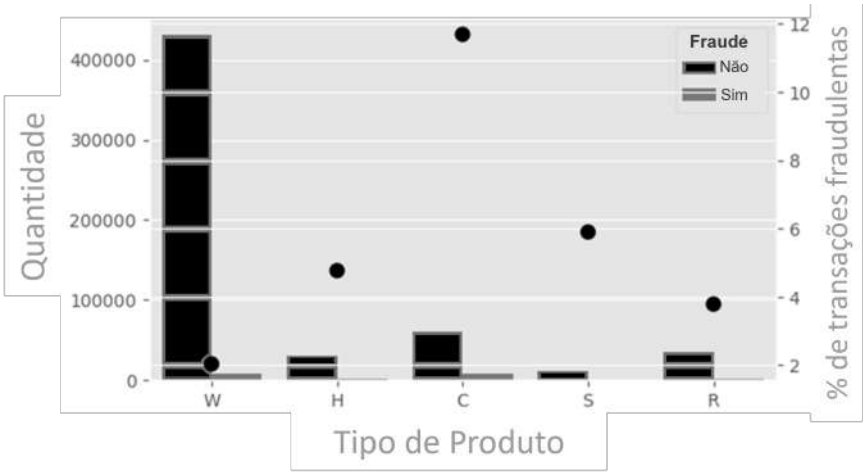


Figura V.1: Quantidade de transações por tipo de produto e variável-alvo. Os círculos indicam os percentuais em termos de transações fraudulentas por tipo de produto.

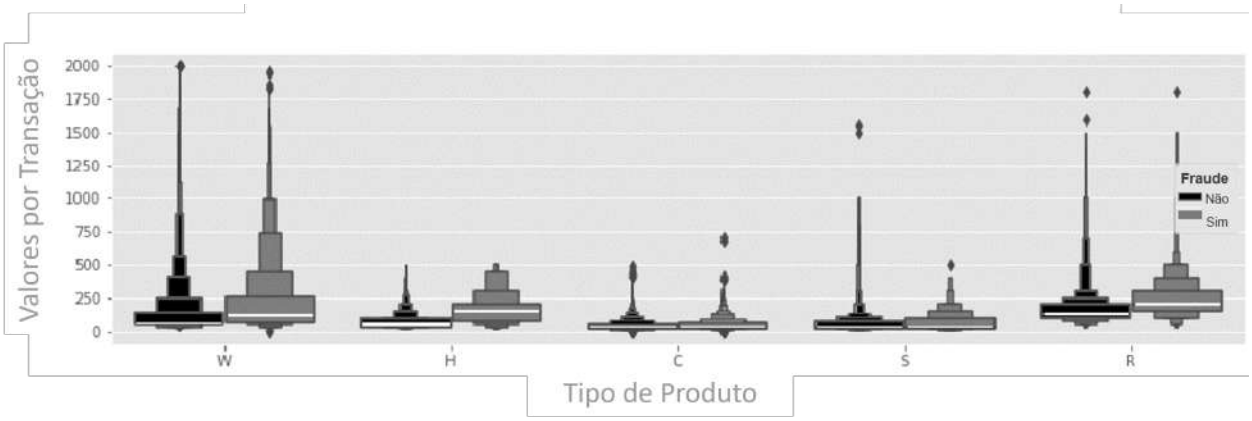


Figura V.2: Boxplots dos valores por transação (em USD), por tipo de Produto e variável-alvo. Produtos do tipo "C" apresentam maior risco de fraude, apesar dos valores maiores por transação dos outros tipos de produto.

Explorando a variável que representa as bandeiras de crédito, entende-se que as maiores proporções de fraude são observadas em cartões "Discover", tanto em termos de quantidade de transações, conforme demonstrado na Figura V.3 (com cerca de 8%), quanto em termos de valor por transação (com cerca de 10%), apesar de serem responsáveis por uma pequena parcela das transações totais. Essa bandeira também apresenta maiores médias de valores por transação, conforme mostrado na Figura V.4.

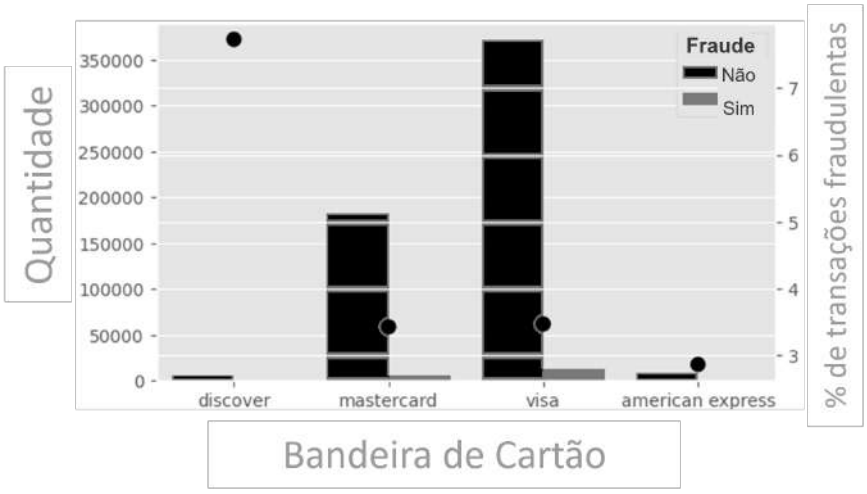


Figura V.3: Quantidade de transações por bandeira e variável-alvo. Os círculos indicam os percentuais em termos de transações fraudulentas por bandeira.

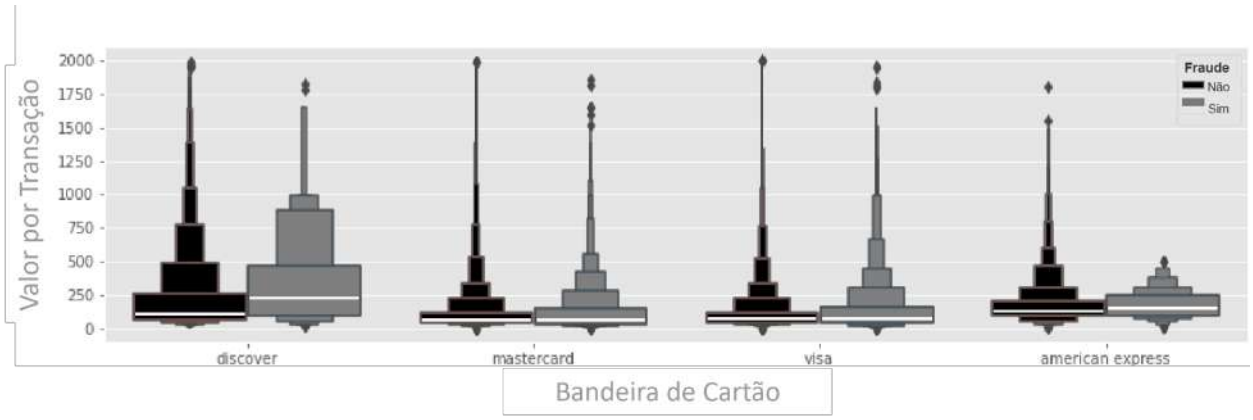


Figura V.4: Boxplots dos valores por transação (em USD), por tipo de produto e variável-alvo. A bandeira "Discover" apresenta maiores médias de valores por transação.

Ainda com relação aos cartões de crédito, conforme pode ser visto na Figura V.5, a categoria de pagamento de crédito apresenta maior proporção de fraude, tanto em número de transações, com cerca de 5%, quanto em valor por transação, com cerca de 6%. Essa proporção apresenta maiores médias de valores por transação, conforme mostrado na Figura V.6.

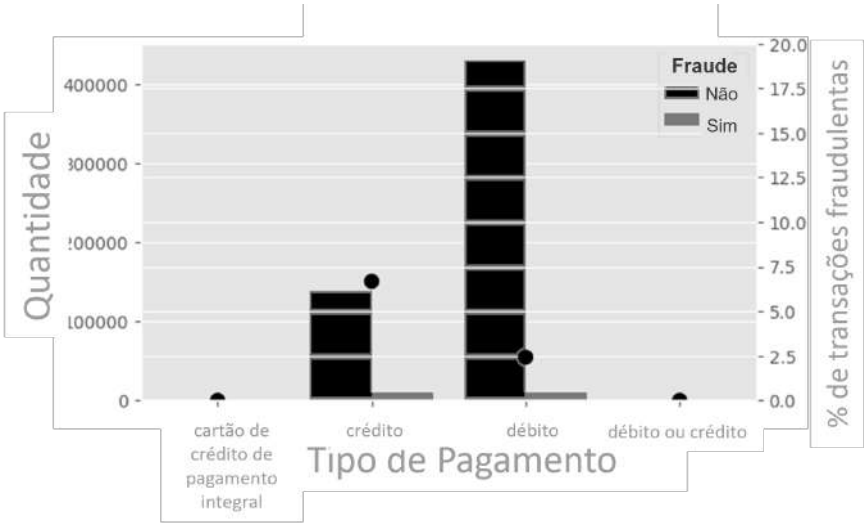


Figura V.5: Quantidade de transações por tipo de pagamento e variável-alvo. Os círculos indicam os percentuais em termos de transações fraudulentas por tipo de pagamento.

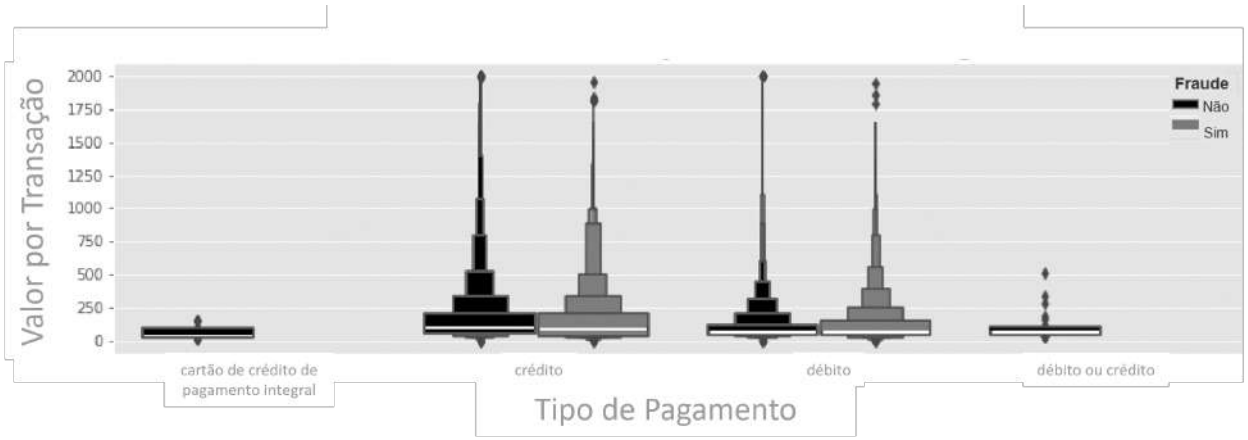


Figura V.6: Boxplots dos valores por transação (em USD), por tipo de pagamento e variável-alvo. "Crédito" apresenta maiores médias de valores por transação.

Finalmente, sobre o comportamento das fraudes em termos de dias no mês, dias da semana e hora, a distribuição de dias no mês mostra que a frequência de transações é relativamente constante ao longo do mês. No entanto, observa-se quedas maiores de número e valores por transação no final do mês, conforme Figuras V.7 e V.8, respectivamente, o que contrasta com o comportamento das taxas de fraude.

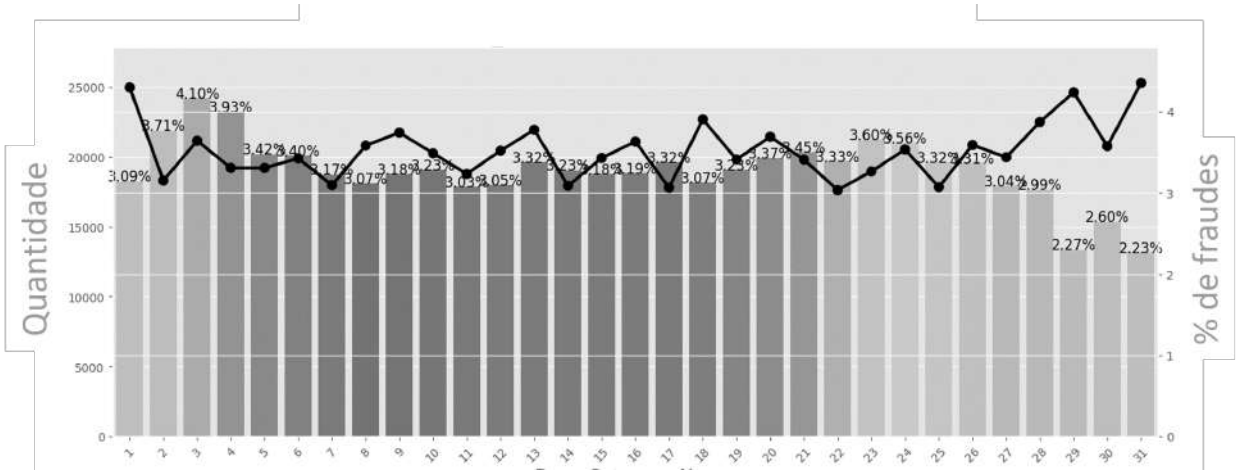


Figura V.7: Total de transações por dia no mês e percentual de fraude. Os percentuais em cima das barras são a representatividade delas em relação ao total de transações no mês.

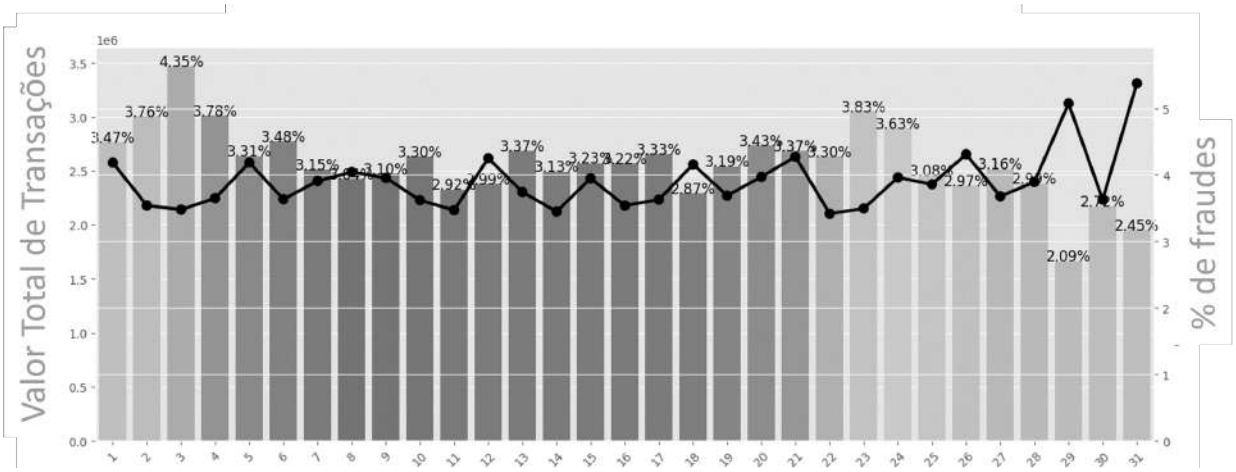


Figura V.8: Total de transações (em milhões de USD) por dia no mês e percentual de fraude. Os percentuais em cima das barras são a representatividade delas em relação ao valor total no mês.

Já com relação à distribuição em dias da semana, os finais de semana - representandos pelos números 5 (sexta-feira), 6 (sábado) e 0 (domingo) - apresentam maiores riscos tanto em termos de transações quanto em termos de valores de fraudes, conforme visto nas Figuras V.9 e V.10, respectivamente. Sobre a distribuição ao longo das horas de um dia, há uma clara correlação negativa entre volume e valores de transações e proporção de fraudes entre 4h e 11h da manhã, conforme visto nas Figuras V.11 e V.12.

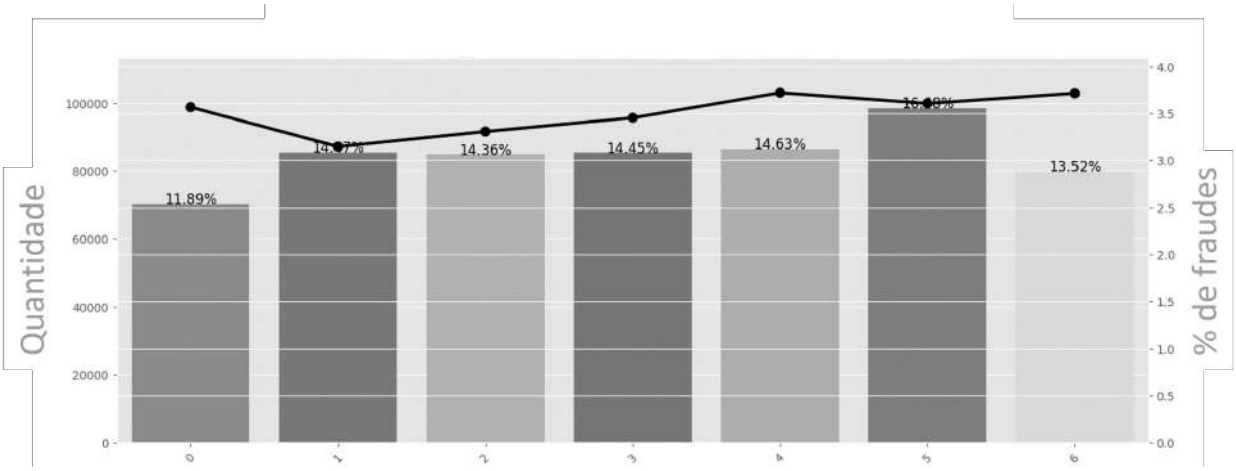


Figura V.9: Total de transações por dia da semana no mês e percentual de fraude. Os percentuais em cima das barras são a representatividade delas em relação ao total de transações na semana.

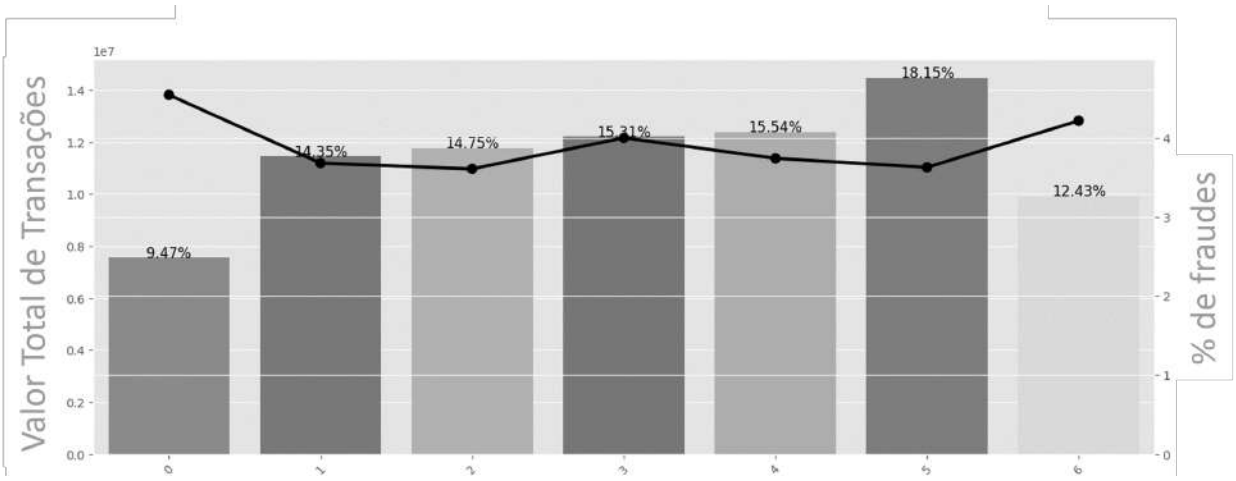


Figura V.10: Total de transações (em bilhões de USD) por dia da semana no mês e percentual de fraude. Os percentuais em cima das barras são a representatividade delas em relação ao valor total na semana.

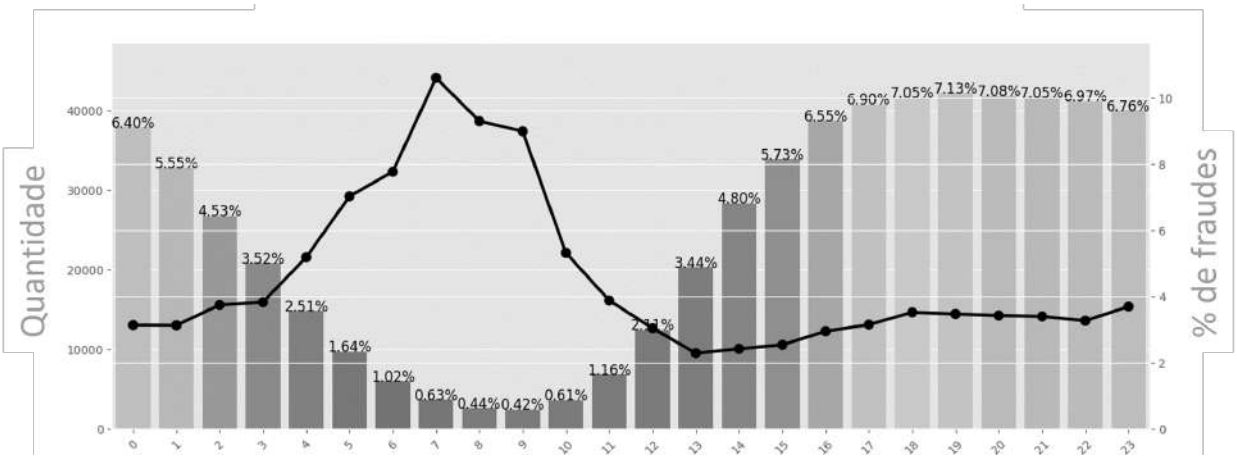


Figura V.11: Total de transações por hora no mês e percentual de fraude. Os percentuais em cima das barras são a representatividade delas em relação ao total de transações no dia.

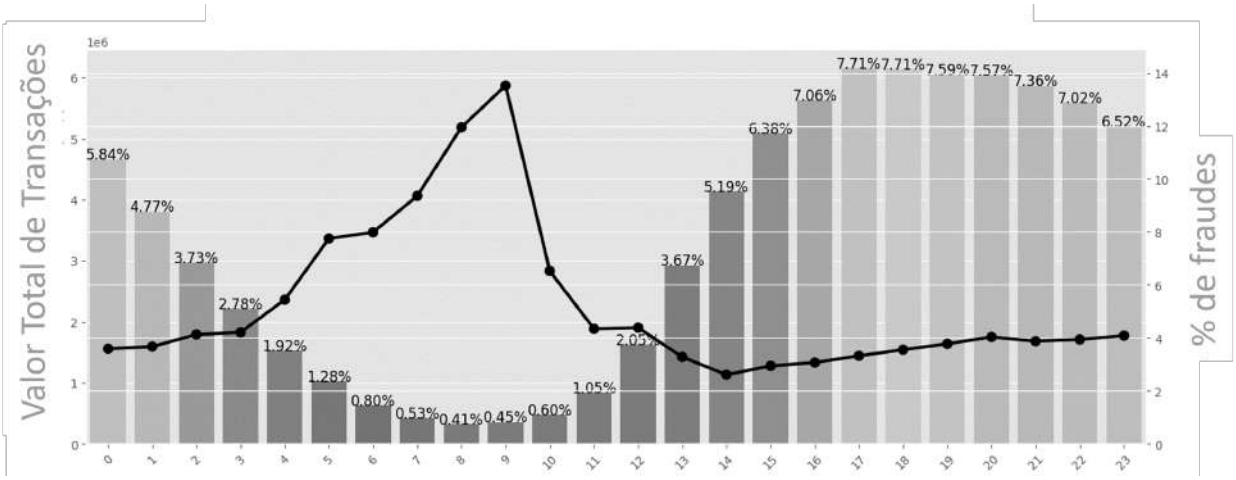


Figura V.12: Total de transações (em milhões de USD) por hora no mês e percentual de fraude. Os percentuais em cima das barras são a representatividade delas em relação ao valor total no dia.

V.3 Análise de Experimentos

Essa seção busca documentar o histórico e evolução de análises realizadas com as diferentes técnicas de seleção de atributos, balanceamento de classes e classificadores discutidas ao longo desse trabalho. Na Seção V.3.1, avaliam-se combinações de técnicas como QQ e ANOVA para seleção de atributos, e métodos de subamostragem e superamostragem, como SMOTE e ROS. Posteriormente, na Seção V.3.2, implementa-se a codificação *one-hot*, estratificação de dados dos conjuntos de treino e teste e validação cruzada durante o treinamento do modelo. Por fim, a Seção V.3.3 retrata a introdução de limiares ótimos de classificação e seu reflexo no desempenho. É importante notar que as Seções V.3.1, V.3.2 e V.3.3 detalham os experimentos iniciais e intermediários realizados exclusivamente com o conjunto de dados IEEE-CIS, que serviu como base para o desenvolvimento do *pipeline* metodológico.

V.3.1 Análise de Técnicas de Seleção de Atributos, de Balanceamento de Classes e Classificadores

Em uma primeira análise do conjunto de dados IEEE-CIS, foram utilizadas: (i) as técnicas de seleção de atributos QQ, ANOVA e AC, em que se avaliaram os 20 e 50 atributos de maior relevância para essas técnicas, além do uso de todos os 88 deles (*baseline*) antes da codificação *one-hot*; (ii) as técnicas de balanceamento de classes de subamostragem *resample*, RUS, *Near Miss* e IHT; (iii) as técnicas de balanceamento de classes de superamostragem *resample*, ROS, SMOTE, BSMOTE e ADASYN; e (iv) os classificadores AD, ET, FA, GB, KNN, NB e RL. Importante ressaltar que, para essa fase exploratória, foram avaliados os 20 e 50 atributos de maior relevância para as respectivas técnicas, limiares escolhidos para interpolar e verificar quais faixas de seleção de atributos possuem melhor desempenho dentro dos *pipelines*.

Os melhores resultados em termos de escore F_1 foram alcançados utilizando métodos de superamostragem, especificamente *resample* e ROS, utilizando *baseline*, como podem ser vistos na Tabela V.1. Essas abordagens atingiram valores de escore F_1 superiores a outras técnicas de balanceamento, especialmente quando combinadas com classificadores como ET e AD. BSMOTE também mostrou-se uma alternativa viável. Embora a subamostragem com *Near Miss* tenha mostrado desempenho competitivo em alguns casos, em geral, as técnicas de subamostragem foram superadas pelos métodos de superamostragem.

No que diz respeito à seleção de atributos, o teste de QQ, utilizando os 50 atributos de maior relevância segundo o teste, proporcionou um desempenho superior ao de técnicas como AC e ANOVA, especialmente em conjunto com técnicas de superamostragem e classificadores como ET. Outra estratégia eficaz é o uso de *baseline* do conjunto de dados, que mantém valores de escore F_1 semelhantes ao do teste de QQ. Por outro lado, AC e ANOVA - técnicas de redução de dimensionalidade que selecionam os atributos mais relevantes em uma relação não-linear com a variável-alvo - apresentam desempenhos

Seleção de Atributos	Ranqueamento	Classificador	Escore F_1 por Método de Balanceamento de Classes								
			Sub-Resample	RUS	Near Miss	IHT	Sup-Resample	ROS	SMOTE	BSMOTE	AdaSyn
QQ	top 20	ET	0,750	0,750	0,953	0,703	0,979	0,979	0,951	0,965	0,947
		FA	0,717	0,724	0,899	0,001	0,722	0,723	0,805	0,816	0,751
		AD	0,701	0,695	0,955	0,647	0,951	0,951	0,920	0,934	0,912
		GB	0,744	0,754	0,945	0,344	0,763	0,761	0,843	0,851	0,828
		RL	0,666	0,666	0,917	0,317	0,676	0,674	0,669	0,686	0,645
		KNN	0,667	0,671	0,946	0,529	0,909	0,910	0,897	0,925	0,885
		NB	0,672	0,672	0,913	0,182	0,684	0,684	0,683	0,676	0,688
	top 50	ET	0,793	0,799	0,949	0,789	**0,990**	**0,990**	0,970	0,979	0,969
		FA	0,739	0,742	0,901	0,193	0,741	0,740	0,740	0,740	0,721
		AD	0,728	0,722	0,941	0,687	0,962	0,962	0,942	0,951	0,940
		GB	0,768	0,779	0,953	0,416	0,785	0,784	0,884	0,896	0,885
		RL	0,677	0,680	0,933	0,408	0,684	0,684	0,683	0,689	0,668
		KNN	0,673	0,677	0,945	0,556	0,916	0,916	0,908	0,934	0,899
		NB	0,692	0,692	0,917	0,431	0,703	0,704	0,702	0,705	0,709
ANOVA	top 20	ET	0,680	0,665	0,831	0,464	0,725	0,724	0,822	0,836	0,832
		FA	0,640	0,642	0,763	0,143	0,651	0,653	0,663	0,736	0,777
		AD	0,676	0,660	0,831	0,448	0,725	0,724	0,822	0,837	0,833
		GB	0,688	0,682	0,826	0,081	0,718	0,716	0,786	0,808	0,811
		RL	0,688	0,693	0,791	0,258	0,695	0,696	0,726	0,734	0,717
		KNN	0,640	0,603	0,813	0,452	0,690	0,674	0,785	0,818	0,777
		NB	0,561	0,566	0,793	0,370	0,555	0,555	0,561	0,558	0,502
	top 50	ET	0,769	0,774	0,917	0,695	0,926	0,925	0,947	0,957	0,942
		FA	0,702	0,707	0,897	0,174	0,713	0,712	0,632	0,738	0,771
		AD	0,720	0,729	0,920	0,628	0,910	0,909	0,936	0,945	0,933
		GB	0,758	0,765	0,925	0,081	0,768	0,766	0,845	0,855	0,856
		RL	0,720	0,725	0,907	0,458	0,728	0,726	0,739	0,753	0,726
		KNN	0,699	0,710	0,895	0,496	0,873	0,874	0,863	0,872	0,842
		NB	0,667	0,672	0,903	0,416	0,675	0,674	0,687	0,677	0,663
AC	top 20	ET	0,696	0,683	0,900	0,585	0,880	0,878	0,882	0,903	0,844
		FA	0,674	0,683	0,865	0,160	0,668	0,670	0,712	0,760	0,687
		AD	0,664	0,657	0,900	0,561	0,872	0,869	0,877	0,899	0,841
		GB	0,717	0,720	0,890	0,430	0,722	0,719	0,821	0,848	0,797
		RL	0,622	0,629	0,880	0,331	0,625	0,629	0,661	0,701	0,627
		KNN	0,677	0,662	0,864	0,373	0,828	0,833	0,782	0,798	0,755
		NB	0,558	0,561	0,881	0,374	0,553	0,552	0,559	0,559	0,498
	top 50	ET	0,790	0,792	0,954	0,748	0,993	0,993	0,975	0,982	0,975
		FA	0,658	0,671	0,901	0,280	0,654	0,655	0,715	0,764	0,669
		AD	0,705	0,703	0,933	0,619	0,958	0,959	0,940	0,948	0,937
		GB	0,757	0,762	0,954	0,535	0,771	0,769	0,886	0,899	0,889
		RL	0,648	0,638	0,938	0,079	0,653	0,654	0,635	0,654	0,642
		KNN	0,671	0,674	0,943	0,466	0,916	0,916	0,907	0,930	0,901
		NB	0,610	0,603	0,921	0,402	0,632	0,644	0,652	0,694	0,686
Baseline	Todos	ET	0,811	0,817	0,956	0,795	**0,994**	**0,994**	0,979	0,985	0,980
		FA	0,742	0,746	0,901	0,099	0,741	0,741	0,794	0,768	0,784
		AD	0,728	0,725	0,940	0,686	0,962	0,962	0,949	0,956	0,948
		GB	0,778	0,788	0,957	0,416	0,794	0,792	0,908	0,918	0,913
		RL	0,675	0,681	0,941	0,407	0,688	0,685	0,677	0,696	0,675
		KNN	0,679	0,679	0,944	0,556	0,918	0,918	0,916	0,938	0,910
		NB	0,680	0,683	0,912	0,285	0,687	0,688	0,705	0,705	0,715

Tabela V.1: Desempenho dos modelos (via escore F_1) por configuração experimental (conjunto de dados utilizado: IEEE-CIS)

inferiores e foram desconsiderados nas avaliações posteriores.

Finalmente, entre os classificadores, ET e AD se destacaram como a escolha de melhor desempenho em relação aos outros avaliados (FA, CB, GB, AB, RL, KNN e NB), especialmente quando combinados com técnicas de superamostragem, teste de QQ e o *baseline*.

V.3.2 Análise com Codificação *One-Hot*

Houve mais uma análise de desempenho das técnicas de balanceamento de classes, seleção de atributos e classificadores, utilizando, agora, codificação *one-hot* - que totalizou 163 atributos -, e um maior número de atributos relevantes segundo o QQ. Nesta fase, a seleção de atributos QQ foi avaliada utilizando os 50 e 100 atributos de maior relevância, pois experimentos preliminares demonstraram que esses limiares proporcionavam melhor desempenho do que a seleção de quantidades menores (como 25 e 75 atributos). Como pode ser visto na Tabela V.2, os resultados ainda indicaram que as melhores combinações para maximizar o desempenho, considerando a métrica de escore F_1, estão relacionados principalmente aos métodos de superamostragem, como *resample*, ROS e BSMOTE. Esses métodos, em combinação com ET e AD, atingirem valores de escore F_1 consistentemente altos, com destaque para ET. Dessa forma, os classificadores FA, CB, GB, AB, RL, KNN e NB foram removidos dos experimentos posteriores.

Em relação às técnicas de seleção de atributos, o *baseline* apresentou um desempenho consistentemente superior que os 50 e 100 atributos de maior relevância segundo o teste QQ - especialmente com ET e métodos de superamostragem. No entanto, a configuração de 100 atributos de maior relevância segundo o teste QQ foi mantida em avaliações posteriores para efeito de comparação com a abordagem *baseline*.

V.3.3 Análise com Limiares Ótimos, Estratificação de Dados e Validação Cruzada

Houve, após a codificação *one-hot*, a implementação das técnicas de limiares ótimos, estratificação de dados e validação cruzada. Com a introdução da busca de limiares ótimos como técnica de balanceamento, houve melhoria de desempenho do modelo comparativamente com outras técnicas, como pode ser visto na Tabela V.3. A combinação entre ET, no intervalo de limiares entre 0,20 e 0,35, proporcionou melhores escores F_1 que as técnicas *resample*, SMOTE e ADASYN. Técnicas de subamostragem, como *Near Miss*, ainda exibiram desempenho inferior às outras. Contudo, limiares mais extremos, como 0,05 ou 0,75, apresentaram desempenho reduzido. Vale ressaltar que o cálculo do escore F_1 foi feito a partir da média de 10 *folds* e que calculou-se também o desvio-padrão a partir dessas 10 execuções. Com isso, alguns valores de escore F_1 estão estatisticamente empatados entre si, considerando a margem de erro definida pelo desvio-padrão, o que é sinalizado pelo sublinhado desses valores.

Em termos de seleção de atributos, tanto QQ (100 melhores atributos) quanto *baseline* apresentaram

Seleção de Atributos	**Ranqueamento**	**Classificador**	**Escore F_1 por Método**					
			Near Miss	Sup-Resample	ROS	SMOTE	BSMOTE	AdaSyn
Qui-Quadrado	top 50	ET	0,937	0,997	0,997	0,987	0,989	0,982
		FA	0,885	0,739	0,742	0,775	0,820	0,744
		AD	0,915	0,988	0,988	0,971	0,971	0,961
		GB	0,938	0,792	0,793	0,930	0,929	0,935
		RL	0,884	0,681	0,680	0,687	0,675	0,686
		KNN	0,897	0,972	0,972	0,943	0,960	0,912
		NB	0,846	0,608	0,611	0,685	0,699	0,714
	top 100	ET	0,944	**0,999**	**0,999**	0,990	0,989	0,987
		FA	0,873	0,735	0,744	0,752	0,829	0,752
		AD	0,906	0,988	0,988	0,973	0,973	0,963
		GB	0,936	0,798	0,798	0,943	0,944	0,950
		RL	0,883	0,678	0,677	0,687	0,685	0,686
		KNN	0,879	0,976	0,976	0,951	0,964	0,921
		NB	0,851	0,671	0,675	0,752	0,762	0,752
Baseline	Todos	ET	0,944	**0,999**	**0,999**	0,990	0,989	0,988
		FA	0,864	0,726	0,728	0,767	0,832	0,759
		AD	0,906	0,988	0,989	0,973	0,973	0,964
		GB	0,937	0,799	0,801	0,947	0,948	0,951
		LR	0,882	0,677	0,678	0,687	0,685	0,685
		KNN	0,878	0,976	0,976	0,951	0,964	0,921
		NB	0,850	0,676	0,678	0,755	0,763	0,753

Tabela V.2: Desempenho dos modelos (via escore F_1) por configuração experimental, após codificação *one-hot* (conjunto de dados utilizado: IEEE-CIS)

Seleção de Atributos	Classificador	Balanceamento de Classes	Escore F_1
QQ (top 100)	ET	N/A	0,651 ± 0,002
		Ajuste de limiares (Limiar: 0,05)	0,336 ± 0,002
		Ajuste de limiares (Limiar: 0,10)	0,543 ± 0,002
		Ajuste de limiares (Limiar: 0,15)	0,654 ± 0,002
		Ajuste de limiares (Limiar: 0,20)	0,708 ± 0,003
		Ajuste de limiares (Limiar: 0,25)	0,727 ± 0,002
		Ajuste de limiares (Limiar: 0,30)	**0,729** ± 0,001
		Ajuste de limiares (Limiar: 0,35)	0,717 ± 0,002
		Ajuste de limiares (Limiar: 0,40)	0,701 ± 0,002
		Ajuste de limiares (Limiar: 0,45)	0,679 ± 0,002
		Ajuste de limiares (Limiar: 0,50)	0,654 ± 0,002
		Ajuste de limiares (Limiar: 0,55)	0,626 ± 0,003
		Ajuste de limiares (Limiar: 0,60)	0,595 ± 0,002
		Ajuste de limiares (Limiar: 0,65)	0,560 ± 0,002
		Ajuste de limiares (Limiar: 0,70)	0,520 ± 0,002
		Ajuste de limiares (Limiar: 0,75)	0,474 ± 0,004
		Sub-Resample	0,329 ± 0,001
		RUS	0,329 ± 0,001
		Near Miss	0,070 ± 0,000
		IHT	0,405 ± 0,001
		Sup-Resample	0,644 ± 0,003
		ROS	0,640 ± 0,002
		SMOTE	0,685 ± 0,003
		BSMOTE	0,671 ± 0,002
		AdaSyn	0,680 ± 0,002
	AD	N/A	**0,542** ± 0,002
		Ajuste de limiares (Limiar: 0,05-0,75)	**0,542** ± 0,002
		Sub-Resample	0,191 ± 0,001
		RUS	0,190 ± 0,001
		Near Miss	0,070 ± 0,000
		IHT	0,291 ± 0,001
		Sup-Resample	0,539 ± 0,001
		ROS	0,542 ± 0,002
		SMOTE	0,503 ± 0,003
		BSMOTE	0,461 ± 0,002
		AdaSyn	0,490 ± 0,002
baseline (163 atributos)	ET	N/A	0,549 ± 0,002
		Ajuste de limiares (Limiar: 0,05)	0,341 ± 0,002
		Ajuste de limiares (Limiar: 0,10)	0,544 ± 0,002
		Ajuste de limiares (Limiar: 0,15)	0,654 ± 0,002
		Ajuste de limiares (Limiar: 0,20)	0,709 ± 0,002
		Ajuste de limiares (Limiar: 0,25)	0,730 ± 0,002
		Ajuste de limiares (Limiar: 0,30)	**0,732** ± 0,002
		Ajuste de limiares (Limiar: 0,35)	0,721 ± 0,002
		Ajuste de limiares (Limiar: 0,40)	0,704 ± 0,001
		Ajuste de limiares (Limiar: 0,45)	0,681 ± 0,002
		Ajuste de limiares (Limiar: 0,50)	0,654 ± 0,002
		Ajuste de limiares (Limiar: 0,55)	0,623 ± 0,002
		Ajuste de limiares (Limiar: 0,60)	0,591 ± 0,003
		Ajuste de limiares (Limiar: 0,65)	0,554 ± 0,002
		Ajuste de limiares (Limiar: 0,70)	0,512 ± 0,004
		Ajuste de limiares (Limiar: 0,75)	0,465 ± 0,002
		Sub-Resample	0,324 ± 0,001
		RUS	0,322 ± 0,001
		Near Miss	0,070 ± 0,000
		IHT	0,423 ± 0,002
		Sup-Resample	0,646 ± 0,003
		ROS	0,644 ± 0,002
		SMOTE	0,680 ± 0,003
		BSMOTE	0,672 ± 0,003
		AdaSyn	**0,681** ± 0,002
	AD	N/A	0,541 ± 0,004
		Ajuste de limiares (Limiar: 0,05-0,75)	**0,549** ± 0,002
		Sub-Resample	0,190 ± 0,001
		RUS	0,190 ± 0,001
		Near Miss	0,070 ± 0,000
		IHT	0,293 ± 0,001
		Sup-Resample	0,538 ± 0,001
		ROS	0,539 ± 0,001
		SMOTE	0,480 ± 0,001
		BSMOTE	0,465 ± 0,003
		AdaSyn	0,499 ± 0,003

Tabela V.3: Desempenho dos modelos (via escore F_1) por configuração experimental, após implementação de limiares ótimos, estratificação de dados e validação cruzada (conjunto de dados utilizado: IEEE-CIS)

desempenho semelhante, mas o último se destaca por alcançar resultados melhores com superamostragem, especialmente quando combinado com ET. Por fim, ET mostrou-se consistentemente superior em relação à AD em quase todas as combinações de técnicas de balanceamento e seleção de atributos.

V.4 Resultados Alcançados

Os experimentos finais consistiram na utilização de: (i) QQ e *baseline* como técnicas de seleção de atributos; (ii) *resample*, RUS, *Near Miss* e IHT como técnicas de balanceamento de classes de subamostragem; (iii) *resample*, ROS, SMOTE, BSMOTE e ADASYN como técnicas de balanceamento de classes de superamostragem e (iv) ET e AD como classificadores. Além da utilização do conjunto de dados IEEE-CIS, também utilizou-se o conjunto de dados MLG-ULB como base de validação para os resultados obtidos no primeiro conjunto. É fundamental enfatizar que, para manter o rigor metodológico e evitar o vazamento de dados, a base de teste foi previamente isolada e não sofreu intervenção de técnicas de balanceamento ou seleção de atributos. Consequentemente, todas as técnicas foram aplicadas exclusivamente sobre a base de treino, e os escores F_1 apresentados a seguir são resultados da avaliação no conjunto de teste original. Finalmente, vale ressaltar que, para o conjunto IEEE-CIS, as técnicas de superamostragem chegaram a uma proporção de 50/50 (570 mil / 570 mil) no balanceamento, enquanto as técnicas de subamostragem chegaram a uma proporção de 90/10 (230 mil / 20 mil). O ajuste de limiares não gerou balanceamento, ficando com a proporção de 570 mil transações genuínas e 20 mil fraudulentas.

A versão final do *pipeline*, bem como a Tabela V.4, apresentam apenas a combinação de métodos de seleção de atributos (QQ e *baseline*), ajustes de limiares (0,25 e 0,55), técnicas de superamostragem, técnicas de subamostragem e classificadores (AD e ET) com melhores desempenhos. Também são apresentados os resultados em que nenhuma técnica de balanceamento é implementado (N/A) para efeitos de comparação. É importante ressaltar que cada variação da tabela foi executada dez vezes para cálculo da média e do desvio padrão.

No caso do conjunto de dados IEEE-CIS, observa-se que QQ (100 maiores atributos) e *baseline* apresentam desempenhos próximos, mas com esse último tendo uma ligeira vantagem quando combinado com o ajuste de limiares e técnicas de superamostragem. A rigor, o ajuste de limiar a 0,30, quando combinado com a abordagem *baseline* e o classificador ET, apresentou o melhor desempenho em termos de escore F_1 - válido também para a combinação com QQ, ligeiramente inferior ao primeiro caso. Em contrapartida, combinações com AD se mostraram limitadas em relação a ET.

O comportamento é semelhante com relação ao conjunto de dados MLG-ULB, com ET, superamostragem e ajustes de limiares apresentando desempenho superior, em detrimento de AD, com desempenho inferior. Mais especificamente, BSMOTE com ET e seleção dos 20 melhores atributos pelo QQ alcançou esse resultado superior, em conjunto com ET e ajuste de limiares no valor de 0,25. A aborda-

gem com o IHT também apresentou, em média, um desempenho inferior, mas, no geral e comparativamente, houve uma diferença em termos de desempenho da opção sem técnicas de balanceamento (N/A): no caso do conjunto de dados MLG-ULB, o valor do escore F_1 foi comparável ao do ajuste de limiar e da técnica de superamostragem BSMOTE.

Conjunto de dados	Seleção de Atributos	Classificador	Balanceamento de Classes	Escore F_1
IEEE-CIS	QQ (top 100)	ET	N/A	0,651 ± 0,002
			Ajuste de limiares (Limiar: 0,30)	**0,729** ± 0,001
			IHT	0,405 ± 0,001
			SMOTE	0,685 ± 0,003
		AD	N/A	0,542 ± 0,002
			Ajuste de limiares (Limiar: 0,55)	0,542 ± 0,002
			IHT	0,291 ± 0,001
			ROS	0,542 ± 0,002
	baseline (163 atributos)	ET	N/A	0,549 ± 0,002
			Ajuste de limiares (Limiar: 0,30)	**0,732** ± 0,002
			IHT	0,423 ± 0,002
			AdaSyn	0,681 ± 0,002
		AD	N/A	0,541 ± 0,004
			Ajuste de limiares (Limiar: 0,55)	0,549 ± 0,002
			IHT	0,293 ± 0,001
			ROS	0,539 ± 0,001
Worldline-MLG	QQ (top 20)	ET	N/A	0,830 ± 0,005
			Ajuste de limiares (Limiar: 0,40)	0,839 ± 0,009
			IHT	0,779 ± 0,009
			BSMOTE	**0,860** ± 0,008
		AD	N/A	0,690 ± 0,022
			Ajuste de limiares (Limiar: 0,55)	0,695 ± 0,014
			IHT	0,556 ± 0,014
			ROS	0,696 ± 0,022
	baseline (30 atributos)	ET	N/A	0,700 ± 0,021
			Ajuste de limiares (Limiar: 0,25)	**0,862** ± 0,006
			IHT	0,760 ± 0,006
			SMOTE	0,843 ± 0,008
		AD	N/A	0,695 ± 0,018
			Ajuste de limiares (Limiar: 0,55)	0,710 ± 0,016
			IHT	0,559 ± 0,011
			ROS	0,723 ± 0,012

Tabela V.4: Síntese dos resultados que alcançaram melhores desempenhos de escore F_1

Esta seção apresentou a síntese dos principais achados do trabalho, em alinhamento com o objetivo central levantado no Capítulo I: avaliar o desempenho das combinações de técnicas em cenários de desbalanceamento. De forma geral, os experimentos finais validaram a eficácia do *pipeline* proposto e confirmaram um ganho de desempenho significativo (medido pelo escore F_1) quando são implementadas técnicas de balanceamento de dados e otimização de limiares de classificação. Observou-se a superioridade das abordagens de superamostragem e de ajuste de limiares em relação à subamostragem. As combinações mais eficazes envolveram o classificador ET, que demonstrou desempenho consistentemente superior a AD na maioria das configurações.

Capítulo VI Conclusões

Os resultados deste trabalho destacam a eficácia das técnicas de balanceamento de dados na detecção de fraudes em transações online de cartão de crédito, evidenciando que estratégias como superamostragem e ajustes de limiares proporcionam melhorias substanciais na identificação de transações fraudulentas. A análise realizada mostrou que a aplicação dessas técnicas, aliada à otimização de classificadores, possibilitou um aumento no desempenho dos modelos, medido principalmente pela métrica escore F_1.

Tanto a seleção de atributos com base no teste QQ, considerando os 100 atributos mais relevantes, quanto a utilização de todos os atributos (*baseline*) apresentaram desempenho superior em relação às demais técnicas de seleção. Em relação aos classificadores, ET demonstrou desempenho superior em comparação a AD na maioria das configurações testadas. A validação do *pipeline* em um conjunto de dados distinto do de treinamento e o foco em métricas de desempenho apropriadas para dados desbalanceados (escore F_1) reforçam a relevância dos resultados obtidos.

Em relação às técnicas de balanceamento, a superamostragem com SMOTE e BSMOTE e o ajuste de limiares de decisão contribuíram de forma decisiva para o aprimoramento das métricas de desempenho. Nesse último caso, ajustes específicos nos limiares podem maximizar o escore F_1, balanceando a sensibilidade e a especificidade dos modelos. Por outro lado, métodos de subamostragem apresentaram limitações devido à possível perda de informações valiosas das classes majoritárias.

As análises realizadas neste trabalho contribuíram para um entendimento mais preciso sobre o impacto das técnicas de balanceamento na detecção de fraudes em transações de cartão de crédito, fornecendo subsídios relevantes para a escolha de estratégias em cenários específicos. Como parte do esforço de disseminação científica, foi publicado o artigo "Avaliação de Técnicas de Balanceamento de Dados na Detecção de Fraude em Transações Online de Cartão de Crédito" no Simpósio Brasileiro de Banco de Dados (SBBD) de 2024[1].

Como desdobramentos futuros, sugere-se explorar abordagens híbridas que integrem ajustes de limiares e técnicas de balanceamento, além de investigar o desempenho de classificadores por meio da otimização de seus hiperparâmetros. Também é relevante considerar melhorias nas técnicas de balanceamento avaliadas ou, ainda, desenvolver novos métodos que ampliem a eficácia no tratamento de conjuntos de dados desbalanceados. Adicionalmente, pode-se avaliar a aplicação de técnicas de apren-

[1] `https://sol.sbc.org.br/index.php/sbbd/article/view/30734`

dizado profundo, como (GAN), que podem ser exploradas em conjunto com as estratégias de balanceamento para aprofundar a detecção de padrões complexos — embora sua eficácia nesse contexto específico ainda requeira validação empírica adicional.

Referências Bibliográficas

J. Anita Smiles and T. Kamalakannan. Data mining based hybrid latent representation induced ensemble model towards fraud prediction. *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 376–382, 2020. 1

Vinicius Facco Rodrigues, Lucas Micol Policarpo, Diórgenes Eugênio da Silveira, Rodrigo da Rosa Righi, Cristiano André da Costa, Jorge Luis Victória Barbosa, Rodolfo Stoffel Antunes, Rodrigo Scorsatto, and Tanuj Arcot. Fraud detection and prevention in e-commerce: A systematic literature review. *Electronic Commerce Research and Applications*, 56:101207, 2022. ISSN 1567-4223. doi: https://doi.org/10.1016/j.elerap.2022.101207. URL https://www.sciencedirect.com/science/article/pii/S1567422322000904. 1, 4

Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, and Björn Ottersten. Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, 51:134–142, 2016. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2015.12.030. URL https://www.sciencedirect.com/science/article/pii/S0957417415008386. 1, 2, 14

Utkarsh Porwal and Smruthi Mukund. Credit card fraud detection in e-commerce. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 280–287, 2019. doi: 10.1109/TrustCom/BigDataSE.2019.00045. 1

Alexander Artikis, Nikos Katzouris, Ivo Correia, Chris Baber, Natan Morar, Inna Skarbovsky, Fabiana Fournier, and Georgios Paliouras. A prototype for credit card fraud management: Industry paper. In *Proceedings of the 11th ACM International Conference on Distributed and Event-Based Systems*, DEBS '17, page 249–260, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350655. 1

Siddhartha Bhattacharyya, Sanjeev Jha, Kurian K. Tharakunnel, and J. Christopher Westland. Data mining for credit card fraud: A comparative study. *Decis. Support Syst.*, 50:602–613, 2011. 1, 4, 10, 14

Yixuan Zhang, Jialiang Tong, Ziyi Wang, and Fengqiang Gao. Customer transaction fraud detection using xgboost model. In *2020 International Conference on Computer Engineering and Application (ICCEA)*, pages 554–558, 2020. 1

Abdulla Muaz, Manoj Jayabalan, and Vinesh Thiruchelvam. A comparison of data sampling techniques for credit card fraud detection. *International Journal of Advanced Computer Science and Applications*, 11(6), 2020. doi: 10.14569/IJACSA.2020.0110660. URL http://dx.doi.org/10.14569/IJACSA.2020.0110660. 2, 7, 11, 22, 23, 24, 27

A. Ali, Siti Mariyam Hj. Shamsuddin, and Anca L. Ralescu. Classification with class imbalance problem: A review. In *Soft Computing Models in Industrial and Environmental Applications*, 2015. 2

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239. 2, 7, 8

YANMIN SUN, ANDREW K. C. WONG, and MOHAMED S. KAMEL. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719, 2009. doi: 10.1142/S0218001409007326. URL https://doi.org/10.1142/S0218001409007326. 2, 7

Sotiris Kotsiantis, D. Kanellopoulos, and P. Pintelas. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30:25–36, 11 2005. 2, 8

S Md. S Askari and Md. Anwar Hussain. Credit card fraud detection using fuzzy id3. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 446–452, 2017. doi: 10.1109/CCAA.2017.8229897. 4

Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68:90–113, 2016. ISSN 1084-8045. doi: https://doi.org/10.1016/j.jnca.2016.04.007. URL https://www.sciencedirect.com/science/article/pii/S1084804516300571. 4, 5, 7, 10

Niek Tax, Kees Jan de Vries, Mathijs de Jong, Nikoleta Dosoula, Bram van den Akker, Jon Smith, Olivier Thuong, and Lucas Bernardi. Machine learning for fraud detection in e-commerce: A research agenda. In Gang Wang, Arridhana Ciptadi, and Ali Ahmadzadeh, editors, *Deployable Machine Learning for Security Defense*, pages 30–54, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87839-9. 5, 6

Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, 8, 01 2022. doi: 10.1007/s40747-021-00637-x. 5

Yeturu Jahnavi, Poongothai Elango, S. P. Raja, and P. Nagendra Kumar. A novel ensemble stacking classification of genetic variations using machine learning algorithms. *International Journal of Image and Graphics*, 23(02):2350015, 2023. doi: 10.1142/S0219467823500158. URL https://doi.org/10.1142/S0219467823500158. 6

Jeyabharathy Sadaiyandi, Padmapriya Arumugam, Arun Kumar Sangaiah, and Chao Zhang. Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset. *Electronics*, 12(21), 2023. ISSN 2079-9292. doi: 10.3390/electronics12214423. URL https://www.mdpi.com/2079-9292/12/21/4423. 6

P. Kathiravan, P. Shanmugavadivu, and R. Saranya. Mitigating imbalanced data in online social networks using stratified k-means sampling. In *2023 8th International Conference on Business and Industrial Research (ICBIR)*, pages 883–888, 2023. doi: 10.1109/ICBIR57571.2023.10147677. 6

Noor Saleh Alfaiz and Suliman Mohamed Fati. Enhanced credit card fraud detection model using machine learning. *Electronics*, 11(4), 2022. ISSN 2079-9292. doi: 10.3390/electronics11040662. URL https://www.mdpi.com/2079-9292/11/4/662. 7, 8, 10, 11, 12, 13, 14, 15

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 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)*, 42(4): 463–484, 2012. doi: 10.1109/TSMCC.2011.2161285. 8

Spelmen Vimalraj and Porkodi Dr.R. A review on handling imbalanced data. 12 2018. 8

Ronaldo Cristiano Prati, Gustavo Enrique de Almeida Prado Alves Batista, and Maria Carolina Monard. Data mining with imbalanced class distributions: concepts and methods. In *Indian International Conference on Artificial Intelligence - IICAI*. IICAI, 2009. 8, 9

Khan Md. Hasib, Md. Sadiq Iqbal, Faisal Muhammad Shah, Jubayer Al Mahmud, Mahmudul Hasan Popel, Md. Imran Hossain Showrov, Shakil Ahmed, and Obaidur Rahman. A survey of methods for managing the classification and solution of data imbalance problem. *Journal of Computer Science*, 16(11): 1546–1557, Nov 2020. doi: 10.3844/jcssp.2020.1546.1557. URL https://thescipub.com/abstract/jcssp.2020.1546.1557. 8

Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv.*, 52(4), aug 2019. ISSN 0360-0300. doi: 10.1145/3343440. URL https://doi.org/10.1145/3343440. 8

Tuong Le, Le Hoang Son, Minh Thanh Vo, Mi Young Lee, and Sung Wook Baik. A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset. *Symmetry*, 10(7), 2018. ISSN 2073-8994. doi: 10.3390/sym10070250. URL https://www.mdpi.com/2073-8994/10/7/250. 9

Samaneh Sorournejad, Zahra Zojaji, Reza Atani, and Amir Monadjemi. A survey of credit card fraud detection techniques: Data and technique oriented perspective, 19 Nov 2016 2016. 9, 10, 13

Kavya Gupta, Kirtivardhan Singh, Gaurav Vikram Singh, Mohd. Hassan, Himani, and Upasana Sharma. Machine learning based credit card fraud detection - a review. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 362–368, 2022. doi: 10.1109/ICAAIC53929.2022.9792653. 9, 10

Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, and Gianluca Bontempi. Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, 557:317–331, 2021. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2019.05.042. URL https://www.sciencedirect.com/science/article/pii/S0020025519304451. 9

Thanh Cong Tran and Tran Khanh Dang. Machine learning for prediction of imbalanced data: Credit fraud detection. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–7, 2021. doi: 10.1109/IMCOM51814.2021.9377352. 10, 11, 12, 13

Jonathan Kwaku Afriyie, Kassim Tawiah, Wilhemina Adoma Pels, Sandra Addai-Henne, Harriet Achiaa Dwamena, Emmanuel Odame Owiredu, Samuel Amening Ayeh, and John Eshun. A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. *Decision Analytics Journal*, 6:100163, 2023. ISSN 2772-6622. doi: https://doi.org/10.1016/j.dajour.2023.100163. URL https://www.sciencedirect.com/science/article/pii/S2772662223000036. 10

Karthika J and A. Senthilselvi. Credit card fraud detection based on ensemble machine learning classifiers. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1604–1610, 2022. doi: 10.1109/ICESC54411.2022.9885649. 11

Nagaraju Devarakonda, Sai Rishitha Mulpuri, Srilakshmi Thanmayi Tripuramallu, and Maddi Sathwika. Amount and location based credit card fraud detection. In *2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–8, 2023. doi: 10.1109/ICECCT56650.2023.10179705. 11, 12

Aditi Singh, Anoushka Singh, Anshul Aggarwal, and Anamika Chauhan. Design and implementation of different machine learning algorithms for credit card fraud detection. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–6, 2022. doi: 10.1109/ICECCME55909.2022.9988588. 11

Mohammed Baker, Zuhair Mahmood, and Ehab Shaker. Ensemble learning with supervised machine learning models to predict credit card fraud transactions. *Revue D Intelligence Artificielle*, 34:509–518, 08 2022. doi: 10.18280/ria.360401. 11, 12

Petr Mrozek, John Panneerselvam, and Ovidiu Bagdasar. Efficient resampling for fraud detection during anonymised credit card transactions with unbalanced datasets. In *2020 IEEE/ACM 13th International*

Conference on Utility and Cloud Computing (UCC), pages 426–433, 2020. doi: 10.1109/UCC48980.2020.00067. 12

Waleed Hilal, S. Andrew Gadsden, and John Yawney. Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*, 193:116429, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.116429. URL https://www.sciencedirect.com/science/article/pii/S0957417421017164. 12, 13, 14

Nur Farhana Hordri, Siti Sophiayati Yuhaniz, Nurulhuda Firdaus Mohd Azmi, and Siti Mariyam Shamsuddin. Handling class imbalance in credit card fraud using resampling methods. *International Journal of Advanced Computer Science and Applications*, 9(11), 2018. doi: 10.14569/IJACSA.2018.091155. URL http://dx.doi.org/10.14569/IJACSA.2018.091155. 13

Ranjeet Kumar Ranjan Amit Singh and Abhishek Tiwari. Credit card fraud detection under extreme imbalanced data: A comparative study of data-level algorithms. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(4):571–598, 2022. doi: 10.1080/0952813X.2021.1907795. URL https://doi.org/10.1080/0952813X.2021.1907795. 14

Damanpreet Kaur, Anjali Saini, and Deepti Gupta. Credit card fraud detection using machine learning, deep learning, and ensemble of the both. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 484–489, 2022. doi: 10.1109/PDGC56933.2022.10053175. 15

Xiao Liu, Kuan Yan, Levent Burak Kara, and Zhenguo Nie. Ccfd-net: A novel deep learning model for credit card fraud detection. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 9–16, 2021. doi: 10.1109/IRI51335.2021.00008. 15

Joffrey Leevy, Justin Johnson, John Hancock, and Taghi Khoshgoftaar. Threshold optimization and random undersampling for imbalanced credit card data. *Journal of Big Data*, 10, 05 2023. doi: 10.1186/s40537-023-00738-z. 15, 16, 21, 23, 27

Justin M. Johnson and Taghi M Khoshgoftaar. Deep learning and thresholding with class-imbalanced big data. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 755–762, 2019. doi: 10.1109/ICMLA.2019.00134. 15, 16

D. Padma Prabha and C. Victoria Priscilla. Estimation of optimal threshold shifting to handle class imbalance in credit card fraud detection using machine learning techniques. In *American Institute of Physics Conference Series*, volume 2802 of *American Institute of Physics Conference Series*, page 120014, January 2024. doi: 10.1063/5.0182386. 15, 16, 19, 21, 27

Angela Carrera-Rivera, William Ochoa, Felix Larrinaga, and Ganix Lasa. How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9:101895, 2022. ISSN 2215-0161. doi: https://doi.org/10.1016/j.mex.2022.101895. URL https://www.sciencedirect.com/science/article/pii/S2215016122002746. 17

Emilija Strelcenia and Simant Prakoonwit. A new gan-based data augmentation method for handling class imbalance in credit card fraud detection. In *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 627–634, 2023. doi: 10.1109/SPIN57001.2023.10116543. 18, 23, 24, 25, 28

Anish Mahajan, Vivek Singh Baghel, and Ramkumar Jayaraman. Credit card fraud detection using logistic regression with imbalanced dataset. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 339–342, 2023. 18, 23, 24, 25, 27

Prerak Parekh, Chirag Rana, Kunal Nalawade, and Surekha Dholay. Credit card fraud detection with resampling techniques. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7, 2021. doi: 10.1109/ICCCNT51525.2021.9579915. 18, 23, 25, 27

Yuxin Wang. An ensemble learning imbalanced data classification method based on sample combination optimization. *Journal of Physics: Conference Series*, 1284(1):012035, aug 2019. doi: 10.1088/1742-6596/1284/1/012035. URL https://dx.doi.org/10.1088/1742-6596/1284/1/012035. 18, 23, 24, 25, 28

Ziyan Zhang and Shuyan Huang. Credit card fraud detection via deep learning method using data balance tools. In *2020 International Conference on Computer Science and Management Technology (ICCSMT)*, pages 133–137, 2020. doi: 10.1109/ICCSMT51754.2020.00033. 18, 19, 23, 24, 25, 28

Janet B, Joshua Arul Kumar R, and Didugu Phani Sai Ganesh. Credit card fraud detection with unbalanced real and synthetic dataset using machine learning models. In *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, pages 73–78, 2022. doi: 10.1109/ICESIC53714.2022.9783529. 18, 19, 24, 25, 27

Yalong Xie, Aiping Li, Liqun Gao, Ziniu Liu, and Shan Zhong. A heterogeneous ensemble learning model based on data distribution for credit card fraud detection. *Wirel. Commun. Mob. Comput.*, 2021, January 2021. ISSN 1530-8669. doi: 10.1155/2021/2531210. URL https://doi.org/10.1155/2021/2531210. 18, 19, 23, 24, 28

Patience Chew Yee Cheah, Yue Yang, and Boon Giin Lee. Enhancing financial fraud detection through addressing class imbalance using hybrid smote-gan techniques. *International Journal of Financial*

Studies, 11(3), 2023. ISSN 2227-7072. doi: 10.3390/ijfs11030110. URL https://www.mdpi.com/2227-7072/11/3/110. 18, 20, 27

Chu Shen, Su-Fang Zhang, Jun-Hai Zhai, Ding-Sheng Luo, and Jun-Fen Chen. Imbalanced data classification based on extreme learning machine autoencoder. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 399–404, 2018. doi: 10.1109/ICMLC.2018.8526934. 18, 24, 25, 28

Taj Sapra, Shubhama, and Shweta Meena. A novel approach to handle imbalanced dataset in machine learning. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–5, 2023. doi: 10.1109/I2CT57861.2023.10126309. 18, 28

Emmanuel Ileberi, Yanxia Sun, and Zenghui Wang. Performance evaluation of machine learning methods for credit card fraud detection using smote and adaboost. *IEEE Access*, 9:165286–165294, 2021. doi: 10.1109/ACCESS.2021.3134330. 19, 22, 23, 24, 27

Wei Li, Cheng-shu Wu, and Su-mei Ruan. Cus-rf-based credit card fraud detection with imbalanced data. *Journal of Risk Analysis and Crisis Response*, 12(3), Sep. 2022. doi: 10.54560/jracr.v12i3.332. URL https://www.jracr.com/index.php/jracr/article/view/332. 19, 27

Ibtissam Benchaji, Samira Douzi, and Bouabid ElOuahidi. Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection. In *2018 2nd Cyber Security in Networking Conference (CSNet)*, pages 1–5, 2018. doi: 10.1109/CSNET.2018.8602972. 19, 20, 24, 25, 27

Val Andrei Fajardo, David Findlay, Charu Jaiswal, Xinshang Yin, Roshanak Houmanfar, Honglei Xie, Jiaxi Liang, Xichen She, and D.B. Emerson. On oversampling imbalanced data with deep conditional generative models. *Expert Systems with Applications*, 169:114463, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.114463. URL https://www.sciencedirect.com/science/article/pii/S0957417420311155. 19, 24, 27

Fengjun Zhang, Guanjun Liu, Zhenchuan Li, Chungang Yan, and Changjun Jiang. Gmm-based undersampling and its application for credit card fraud detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. doi: 10.1109/IJCNN.2019.8852415. 20, 24, 25, 28

Akhil Sethia, Raj Patel, and Purva Raut. Data augmentation using generative models for credit card fraud detection. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–6, 2018. doi: 10.1109/CCAA.2018.8777628. 20, 24, 28

Ibomoiye Domor Mienye and Yanxia Sun. A deep learning ensemble with data resampling for credit card fraud detection. *IEEE Access*, 11:30628–30638, 2023. doi: 10.1109/ACCESS.2023.3262020. 21, 23, 27

C. Victoria Priscilla and D. Padma Prabha. Influence of optimizing xgboost to handle class imbalance in credit card fraud detection. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1309–1315, 2020. doi: 10.1109/ICSSIT48917.2020.9214206. 22, 24, 28

Hitesh Bhagwani, Sonali Agarwal, Ashwini Kodipalli, and Roshan Joy Martis. Targeting class imbalance problem using gan. In *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, pages 318–322, 2021. doi: 10.1109/ICEECCOT52851.2021.9708011. 22, 24, 25, 27

Oguz Ata and Layth Hazim. Comparative analysis of different distributions dataset by using data mining techniques on credit card fraud detection. *Tehnicki Vjesnik*, Tehnički vjesnik 27, 2(2020):618–626, 04 2020. doi: 10.17559/TV-20180427091048. 23, 27

J. Karthika and A. Senthilselvi. An integration of deep learning model with navo minority over-sampling technique to detect the frauds in credit cards. *Multimedia Tools Appl.*, 82(14):21757–21774, January 2023. ISSN 1380-7501. doi: 10.1007/s11042-023-14365-6. URL `https://doi.org/10.1007/s11042-023-14365-6`. 23, 27

Fuad A. Ghaleb, Faisal Saeed, Mohammed Al-Sarem, Sultan Noman Qasem, and Tawfik Al-Hadhrami. Ensemble synthesized minority oversampling-based generative adversarial networks and random forest algorithm for credit card fraud detection. *IEEE Access*, 11:89694–89710, 2023. doi: 10.1109/ACCESS.2023.3306621. 23, 27

T. John Berkmans and S. Karthick. Credit card fraud detection with data sampling. In *2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, pages 1–6, 2022. doi: 10.1109/ICPECTS56089.2022.10046729. 24, 27

B. Venkateswarlu, Kasiraju Poornima, Ramineni Vasavi, and Jalireddygari Venkata Vaishnavi. A study on class imbalance problem using genetic algorithm. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1709–1714, 2022. doi: 10.1109/ICSSIT53264.2022.9716371. 24, 25, 28

Tahani Baabdullah, Amani Alzahrani, and Danda B. Rawat. On the comparative study of prediction accuracy for credit card fraud detection with imbalanced classifications. In *Proceedings of the 2020 Spring Simulation Conference*, SpringSim '20, San Diego, CA, USA, 2020. Society for Computer Simulation International. ISBN 9781713812883. 25, 27

Maram Alamri and Mourad Ykhlef. Hybrid undersampling and oversampling for handling imbalanced credit card data. *IEEE Access*, 12:14050–14060, 2024. doi: 10.1109/ACCESS.2024.3357091. 27, 35

N. Sai Sree Pranavi, T. K. S. S. Sruthi, B. Jahnavi Naga Sirisha, M. S. Nayak, and Venkata Sainath Gupta Thadikemalla. Credit card fraud detection using minority oversampling and random forest technique. In *2022 3rd International Conference for Emerging Technology (INCET)*, pages 1–6, 2022. doi: 10.1109/INCET54531.2022.9824146. 28

Jahnavi Yeturu, Poongothai Elango, S Raja, and P Kumar. A novel ensemble stacking classi¯cation of genetic variations using machine learning algorithms. 01 2023. 32

Juan Laborda and Seyong Ryoo. Feature selection in a credit scoring model. *Mathematics*, 9(7), 2021. doi: 10.3390/math9070746. Cited by: 24; All Open Access, Gold Open Access, Green Open Access. 33

CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

TERMO DE ENTREGA DO ORIGINAL DA DISSERTAÇÃO OU TESE À SECRETARIA

Secretaria dos Programas de Pós-Graduação *Stricto Sensu*

PROGRAMA DE PÓS-GRADUAÇÃO/CURSO	
Programa de Pós-Graduação em:	Mestrado (X)
	Doutorado ( )

Título da Dissertação/Tese: Avaliação de Combinações de Técnicas de Balanceamento de Dados e de Seleção de Atributos na Detecção de Fraude em Transações Online de Cartão de Crédito

Autor: Arthur Silveira de Siqueira Cavalcanti

Orientador: Rafaelli de Carvalho Coutinho; Eduardo Bezerra da Silva

Banca Examinadora:
Rafaelli de Carvalho Coutinho
Eduardo Bezerra da Silva
Eduardo Soares Ogasawara
Ronaldo Ribeiro Goldschmidt

Declaro que o original da dissertação de mestrado/tese de doutorado, anexa, está de acordo com a norma para redação de dissertação/tese da DIPPG.

RAFAELLI DE CARVALHO COUTINHO:12132776717
Assinado de forma digital por RAFAELLI DE CARVALHO COUTINHO:12132776717
Dados: 2025.11.09 12:33:16 -03'00'

Prof. Orientador

CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DE DISSERTAÇÕES E TESES EM FORMATO ELETRÔNICO (ON-LINE)

Secretaria dos Programas de Pós-Graduação *Stricto Sensu*

PROGRAMA DE PÓS-GRADUAÇÃO/CURSO

Programa de Pós-Graduação em: Ciência da Computação	Mestrado (X)
	Doutorado ( )

DADOS PESSOAIS DO ALUNO AUTOR

Nome: Arthur Silveira de Siqueira Cavalcanti					
Identidade: 21.179.713-9		CPF: 128.985.797-00		E-mail: arthur.silveirasc@gmail.com	
Telefones	Residencial:	Comercial:		Celular: (21) 9 9991-3838	
Endereço	Rua/Av: Rua Barão de Ipanema		Nº: 143	Complemento: apto 605	
	Bairro Copacabana	Cidade: Rio de Janeiro	UF: RJ	CEP: 22050-031	País: Brasil

NATUREZA DA OBRA: [X] Dissertação de Mestrado [ ] Tese de Doutorado

TÍTULO DA OBRA: Avaliação de Combinações de Técnicas de Balanceamento de Dados e de Seleção de Atributos na Detecção de Fraude em Transações Online de Cartão de Crédito

(SELECIONAR SOMENTE UMA OPÇÃO)

1 [X] **AUTORIZO** ao Centro Federal Tecnológico Celso Suckow da Fonseca – CEFET/RJ a reproduzir, inclusive em outro formato ou mídia e através de armazenamento permanente ou temporário, bem como a publicar na rede mundial de computadores (Internet), na Biblioteca virtual do CEFET/RJ ou de outros Bancos Públicos de Dissertações e Teses, entendendo-se os termos "reproduzir" e "publicar" conforme definições dos incisos VI e I, respectivamente, do artigo 5º da Lei 9610/98 de 10/02/1998, a obra acima mencionada, sem que me seja devido pagamento a título de direitos autorais, desde que a reprodução e/ou publicação tenham a finalidade exclusiva de uso por quem a consulta, e a título de divulgação da produção acadêmica gerada pela Instituição, a partir desta data, e até que, de minha parte, solicite expressamente o cancelamento da presente autorização.

2 [ ] **NÃO AUTORIZO** a disponibilização completa, on-line.

Motivo:

[ ] Exigência de periódico para não divulgação até a publicação (**exige justificativa, informe o periódico**)
[ ] Não envio por exigência contratual (**exige justificativa, informe nº da patente**)
[ ] Outro (**exige justificativa**)

Justificativa:

Rio de Janeiro, 08 de novembro de 2025.

Assinatura do Autor: gov.br Documento assinado digitalmente ARTHUR SILVEIRA DE SIQUEIRA CAVALCANTI Data: 12/11/2025 10:14:40-0300 Verifique em https://validar.iti.gov.br

Ciência do Orientador: Rafaelli de Carvalho Coutinho (NOME)

RAFAELLI DE CARVALHO COUTINHO:12132776717 Assinado de forma digital por RAFAELLI DE CARVALHO COUTINHO:12132776717 Dados: 2025.11.09 12:33:41 -03'00' (ASSINATURA)