UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA

DE PRODUÇÃO

ABORDAGENS EFICIENTES PARA CLASSIFICAÇÃO BINÁRIA EM BASES DE DADOS EXTREMAMENTE DESBALANCEADAS

Leandro Duarte Pereira

Orientador: Prof. Dr. Pedro Paula Balestrassi

Co-orientador: Dr. Fabrício Alves de Almeida

Itajubá

Agosto de 2025

RESUMO

Lidar com dados extremamente desbalanceados em tarefas de classificação binária representa um desafio recorrente em diversos domínios, pois a baixa prevalência da classe minoritária (<1%) compromete a confiabilidade e o desempenho preditivo dos modelos. Embora a literatura apresente um número expressivo de estudos sobre o desbalanceamento, o cenário de desbalanceamento extremo ainda carece de investigações aprofundadas. Nesse contexto, esta tese desenvolveu duas frentes complementares de pesquisa. Na primeira, foi conduzida uma Revisão Sistemática da Literatura (RSL), seguindo rigoroso protocolo de seleção e qualidade, a partir da qual 22 estudos primários foram analisados em 52 bases de dados. Os resultados indicaram que abordagens combinadas apresentam desempenho superior em diversos cenários, destacando-se técnicas de sobreamostragem (oversampling) associadas a ensembles, em especial a combinação de Floresta Aleatória (*Random Forest – RF*) com métodos derivados da Técnica de Sobreamostragem de Minorias Sintéticas (*Synthetic Minority Oversampling Technique – SMOTE*). Na segunda frente, propõe-se uma abordagem inovadora baseada em *Design* de Experimentos (*DoE*) para geração de conjuntos de dados sintéticos em condições de desequilíbrio extremo. A estrutura permite a manipulação controlada de seis fatores críticos (dimensionalidade, tamanho da amostra, razão de desbalanceamento, tipo de função de resposta, limiar de decisão e variabilidade do erro), possibilitando experimentação sistemática e replicável. Experimentos realizados com *Random Forest* combinado ao *SMOTE* evidenciaram a utilidade da estrutura para analisar o impacto de fatores e interações, sendo identificada, por meio de Análise de Variância (*ANOVA*), a relevância da dimensionalidade e da variabilidade do erro no comportamento do classificador. Assim, os achados da Revisão Sistemática da Literatura e a estrutura experimental proposta contribuem de forma integrada para o avanço do conhecimento e para o desenvolvimento de métodos mais robustos em cenários de classificação binária sob desbalanceamento extremo.

Palavras-chave: Desbalanceamento Extremo de Classes, Classificação Binária, Revisão Sistemática da Literatura (RSL), Geração de Dados Sintéticos, *Design* de Experimentos (DoE).

ABSTRACT

Handling extremely imbalanced data in binary classification tasks represents a recurring challenge across multiple domains, as the very low prevalence of the minority class (<1%) compromises both predictive performance and model reliability. Although the literature presents a considerable number of studies on class imbalance, the scenario of extreme imbalance still requires further in-depth investigation. In this context, this thesis developed two complementary research fronts. First, a Systematic Literature Review (SLR) was conducted following a rigorous protocol of selection and quality criteria, through which 22 primary experimental studies were analyzed across 52 datasets. The results indicated that combined approaches achieve superior performance in several scenarios, with particular emphasis on oversampling techniques associated with ensembles, especially the combination of Random Forest (RF) with methods derived from the Synthetic Minority Oversampling Technique (SMOTE). Second, we propose an innovative approach based on Design of Experiments (DOE) for generating synthetic datasets under extreme class imbalance conditions. The framework enables the controlled manipulation of six critical factors (feature dimensionality, sample size, imbalance ratio, response function type, decision threshold, and error variability), allowing systematic and replicable experimentation. Experiments conducted with Random Forest combined with SMOTE demonstrated the usefulness of the framework in analyzing the impact of main effects and interactions, with Analysis of Variance (ANOVA) identifying the relevance of feature dimensionality and error variability to classifier behavior. Altogether, the findings from the Systematic Literature Review and the proposed experimental framework contribute in an integrated manner to advancing knowledge and fostering the development of more robust methods for binary classification under extreme imbalance scenarios.

***Keywords**: Extremely Class Imbalance, Binary Classification, Systematic Literature Review (SLR), Synthetic Data Generation, Design of Experiments (DoE).*

AGRADECIMENTOS

A Deus, pela vida, pela saúde e pela sabedoria concedida ao longo desta caminhada, sem os quais esta conquista não teria sido possível.

À minha família, pelo amor incondicional, paciência e apoio constante em todos os momentos, sendo fonte de força e inspiração.

À Universidade Federal de Itajubá (UNIFEI), pela oportunidade de realização deste Doutorado e pelo ambiente acadêmico de excelência que possibilitou o desenvolvimento desta pesquisa.

Ao meu orientador e ao meu co-orientador, pela orientação dedicada, pela confiança depositada e pelas valiosas contribuições que enriqueceram este trabalho.

Aos professores do Programa de Pós-Graduação em Engenharia de Produção da UNIFEI, pelos ensinamentos, incentivo e pela base sólida transmitida ao longo da formação.

Aos colegas e amigos de trabalho, pelo companheirismo, pelas trocas de experiências e pelo apoio ao longo desta jornada.

A todos que, de alguma forma, contribuíram para a concretização desta tese, expresso minha mais sincera gratidão.

SUMÁRIO

1 INTRODUÇÃO

Os estudos em Inteligência Artificial ou *Artificial Intelligence* têm crescido consideravelmente nos últimos dez anos. A Aprendizagem de Máquina ou *Machine Learning* é provavelmente o ramo mais popular em Inteligência Artificial até o momento. A maioria dos sistemas que utilizam métodos de Aprendizagem de Máquina os utiliza para realizar análises preditivas (BOKONDA; OUAZZANI-TOUHAMI; SOUISSI, 2020). Essas abordagens podem auxiliar em problemas de tomada de decisão em diversas áreas do conhecimento, como medicina, engenharia, economia, educação, entre outras. Por exemplo, a Aprendizagem de Máquina pode ser usada para diagnosticar doenças, otimizar processos, prever demandas, personalizar serviços, etc. De acordo com um relatório da plataforma global de dados e *Business Intelligence Statista*, o tamanho do mercado global de aplicações que fazem uso de Aprendizagem de Máquina era de cerca de 140 bilhões de dólares americanos em 2021 e deverá crescer para quase dois trilhões de dólares americanos até 2030 (THORMUNDSSON, 2023).

Por meio da análise de grandes volumes de dados, os algoritmos de Aprendizagem de Máquina são capazes de identificar padrões e realizar previsões com alta precisão. Isso permite tomadas de decisão mais embasadas e eficientes, otimizando seus processos e obtendo melhores resultados. Todavia para se obter resultados significativos é necessário um bom entendimento do problema a ser resolvido, além de dados de qualidade e um processo adequado de treinamento dos modelos.

A Classificação é uma parte importante da Aprendizagem de Máquina e consideráveis progressos foram feitos em algoritmos e aplicações relacionados. A Classificação permite que as máquinas identifiquem e organizem dados de acordo com critérios pré-definidos. Com a evolução da tecnologia, aprimoramentos contínuos têm sido feitos para melhorar a precisão e eficiência dos Algoritmos de Classificação, tornando-os cada vez mais importantes para o sucesso de aplicações nas mais diversas áreas (WU; ZHANG; HU, 2020).

No entanto, a Classificação Binária é uma das sub áreas de estudos mais frequentes em problemas de Aprendizagem de Máquina. Com pesquisas e aplicações em vários domínios, tem contribuído para importantes avanços em muitas áreas do conhecimento (CANBEK *et al.*, 2017). Esse tipo de classificação é utilizado para separar dados em duas categorias distintas, o que é extremamente útil em diversos cenários do mundo real. A Classificação Binária é um processo complexo que envolve a seleção cuidadosa de variáveis e algoritmos para obter resultados precisos e confiáveis. É importante ressaltar que

a escolha adequada das variáveis e algoritmos pode afetar significativamente a precisão dos modelos de Classificação Binária.

Entretanto, muitas aplicações do mundo real com grandes volumes de dados envolvem bases de dados desbalanceadas, o que pode representar um desafio relevante para os algoritmos de Aprendizagem de Máquina (ZHU *et al.*, 2022).

De acordo com Datar e Garg (2019), o problema de bases de dados desbalanceadas pode ser entendido pelo desbalanceamento de classes que consiste na dificuldade associada aos dados que possuem uma grande maioria de registros pertencentes a uma única classe (majoritária) em detrimento a outra (minoritária).

O desbalanceamento de classes pode levar a problemas como viés nos resultados, dificuldade em identificar corretamente as classes minoritárias e menor acurácia geral do modelo de classificação.

Neste cenário, lidar com problemas de desbalanceamento de classes tem sido amplamente abordado em várias áreas de pesquisa, como por exemplo em modelos de diagnóstico de falhas em bases de dados industriais (REN, Z. *et al.*, 2022). É importante, portanto, que as abordagens de Aprendizagem de Máquina estejam preparadas para lidar com esse tipo de situação empregando técnicas específicas para lidar com bases de dados desbalanceadas. De tal modo, será possível melhorar a precisão e a confiabilidade dos modelos em contextos reais.

O desbalanceamento das bases de dados pode ser categorizado de acordo com a proporção da Classe Minoritária, como demonstrado na Tabela 1. Este critério de classificação é fundamental para compreender a gravidade do desequilíbrio entre as classes e para orientar a escolha das estratégias de balanceamento mais adequadas. A análise da proporção da Classe Minoritária permite identificar se o desbalanceamento é suave, moderado ou extremo, sendo esta última a categoria de interesse desta pesquisa. Esta abordagem classificatória é essencial para garantir que as técnicas de balanceamento sejam aplicadas de forma precisa e eficaz, visando aprimorar o desempenho dos modelos de aprendizado de máquina e minimizar o viés introduzido pelo desbalanceamento. Dessa forma, a compreensão do nível de desbalanceamento das bases de dados é um passo crucial na busca por soluções que promovam a equidade e a precisão nas análises e previsões realizadas a partir desses dados. O desbalanceamento das bases também pode ser representado pelo termo *Imbalance Ratio* (*IR*), de acordo com a notação de Zhu, T. *et al.* (2022):

$$IR = \frac{(\#majority\ class\ samples)}{(\#minority\ class\ samples)} \quad (1)$$

Tabela 1 - Níveis de desbalanceamento.

Nível de desbalanceamento	Proporção da Classe Minoritária
Suave	20-40% do conjunto de dados
Moderado	1-20% do conjunto de dados
Extremo	<1% do conjunto de dados

Fonte: Google for Developers (2025).

Por exemplo, uma base de dados com 5000 registros, sendo 45 registros pertencentes a classe minoritária e 4955 a classe majoritária, tem uma proporção de 0,9% de desbalanceamento e um IR = 110,11 respectivamente.

Todavia é possível encontrar alguns estudos que abordam dados com desbalanceamento Moderado e Suave como se fossem Extremos. Uma hipótese para este problema é que os níveis de desbalanceamento ainda não estão bem padronizados e difundidos na literatura, o que pode dificultar as buscas por abordagens mais precisas e eficientes para lidar com esse problema.

Neste trabalho, realizamos uma análise e comparamos a literatura existente em relação às abordagens que lidam com este problema em estudos primários para identificar lacunas na pesquisa e sugerir direções para trabalhos futuros. Classificamos as abordagens em cada um dos estudos revisados em: técnicas de pré-processamento, classificadores e *ensembles* e avaliamos o desempenho através dos valores das métricas obtidas pelos experimentos. Na sequência, um experimento planejado e controlado foi conduzido com base nos principais achados da RSL com o objetivo de validar na prática as abordagens previamente levantadas.

A estrutura deste trabalho segue a seguinte organização: O Capítulo 1 introduz o assunto, esclarece os objetivos de pesquisa e seus aspectos relevantes. O Capítulo 2 apresenta a Fundamentação Teórica desenvolvida a partir de um protocolo de Revisão Sistemática da Literatura (RSL). O Capítulo 3 apresenta as aplicações experimentais desenvolvidas com base nos principais achados das RSL e propõe uma abordagem para geração de dados sintéticos. O Capítulo 4 traz uma análise dos resultados e conclui o trabalho resumindo as contribuições da pesquisa e fornecendo sugestões para trabalhos futuros.

1.1 Objetivos

1.1.1 Tese

Muitos dos problemas de classificação do mundo real apresentam características extremamente desbalanceadas em seus objetos de estudo. De tal modo, encontrar quais os aspectos são mais relevantes auxiliará pesquisas futuras a prever eventos com a diminuição de falsos positivos e falsos negativos. Esta tese tem como pressuposto demonstrar que problemas de classificação binária em bases extremamente desbalanceadas necessitam de abordagens adequadas para a obtenção de melhores resultados.

1.1.2 Objetivo Geral

O objetivo geral deste trabalho é consolidar e sintetizar o conhecimento existente sobre classificação binária em bases de dados extremamente desbalanceadas e propor um *framework* experimental baseado em DoE para geração e análise de bases sintéticas. Onde, este *framework* permitirá avaliar de forma controlada os fatores que impactam no desempenho de técnicas de classificação em cenários de dados altamente desbalanceados.

1.1.3 Objetivos Específicos

Com vistas a alcançar o objetivo geral, este trabalho se desdobrará nos seguinte objetivos específicos, que podem ser entendidos como entregas principais deste projeto de pesquisa (tese):

i. Revisar a literatura existente contendo apenas estudos primários sobre classificação binária em bases de dados extremamente desbalanceadas através de um protocolo estruturado de Revisão Sistemática da Literatura (RSL).
ii. Catalogar os achados metodológicos e quantitativos da RSL em um banco de dados.
iii. Projetar e implementar um processo controlado para geração de dados sintéticos extremamente desbalanceados por meio de Planejamento de Experimentos.
iv. Conduzir um experimento guiado pelos resultados RSL (abordagens e métricas) tendo como objeto de estudo as bases de dados sintéticas geradas previamente.

1.1.4 Justificativa

Ao longo da última década muitas pesquisas vêm sendo desenvolvidas sobre o problema de dados desbalanceados. Mais especificamente, relacionado a Classificação Binária, uma busca realizada na plataforma *online ScienceDirect* [1], em Novembro de 2023 e com os termos: ("*imbalanced data" OR "unbalanced data") AND ("binary" AND "classification*"), retornou 5134 artigos em diversas áreas do conhecimento com data de publicação superior a 2013.

A maioria dos artigos buscam analisar, comparar ou desenvolver novas abordagens com o objetivo de melhorar a precisão dos modelos através da avaliação dos resultados obtidos por métricas. De acordo com Raghuwanshi e Shukla (2019), os métodos para lidar com bases desbalanceadas se dividem em três categorias principais: Métodos em Nível de Dados, que visam reduzir o desequilíbrio de classes; Métodos em Nível Algorítmico, que modificam o design do classificador para enfrentar o problema; e Métodos Híbridos, que combinam abordagens de nível de dados e nível algorítmico.

Todavia, quanto maior a escala de dados e maior o desbalanceamento, mais difícil e desafiadora se torna a tarefa de Classificação Binária (REN, J. *et al.*, 2022). A escassez de dados em uma das classes pode resultar em resultados imprecisos e gerar dificuldades na tomada de decisão. Portanto, é crucial desenvolver abordagens eficientes para lidar com esse problema, a fim de se garantir modelos mais precisos, confiáveis e não enviesados.

Embora seja encontrado na literatura um número considerável de estudos sobre o desbalanceamento de dados, é importante ressaltar que uma especialização deste assunto, o Desbalanceamento Extremo, ainda carece de uma exploração mais aprofundada, pois apresenta desafios significativos em sua aplicação.

Este fenômeno, caracterizado pela presença de uma desproporção extrema entre as classes de um conjunto de dados, demanda uma abordagem cautelosa e especializada. Em problemas do mundo real, existem diversos casos em que os dados são extremamente desbalanceados (LU *et al.*, 2019), o que torna a tarefa de aprimorar a precisão dos modelos mais desafiadora.

Em uma busca complementar realizada na plataforma online *ScienceDirect*, com a adição do termo "*extreme*": *("extreme imbalanced data" OR "extreme unbalanced data") AND ("binary" AND "classification"),* somente 10 artigos foram encontrados em variadas áreas do conhecimento e com data de publicação superior a 2013. Alguns destes estudos são aplicados em bases de dados reais e também sintéticas.

[1] ScienceDirect.com, Science, Health and Medical Journals, Full Text Articles and Books. <https://www.sciencedirect.com/>

Conforme evidenciado nos trabalhos de Zhu *et al.* (2022); Yuan *et al.* (2023); Xia *et al*. (2021); Noviyanto e Abdulla (2020); Li, Y. *et al*. (2018) e Fan *et al.* (2021) é possível notar que as abordagens relacionadas a técnicas de pré-processamento de dados têm se destacado nas pesquisas mais recentes. Nesse sentido, a atenção voltada para o pré-processamento de dados demonstra o reconhecimento da relevância de se preparar adequadamente as informações antes de sua utilização em análises e modelagens. Essa tendência reflete não apenas a necessidade de lidar com conjuntos de dados cada vez mais complexos e diversificados, mas também a busca por resultados mais confiáveis e significativos. Dessa forma, as abordagens relacionadas ao pré-processamento de dados constituem um campo de estudo promissor e em constante evolução, com impactos significativos em diversas áreas do conhecimento e da prática profissional.

Outras estratégias bem-sucedidas baseiam-se no emprego de algoritmos de *ensembles*, como demonstrado nos artigos de Zięba *et al.* (2014) e Zhao, S. *et al.* (2020) . A utilização de algoritmos de ensembles permite a combinação de múltiplos modelos, resultando em previsões mais precisas e robustas. Além disso, essa técnica é capaz de lidar com a complexidade inerente a conjuntos de dados de grande escala, proporcionando uma maior generalização e capacidade de adaptação a diferentes cenários. Dessa forma, os algoritmos de ensembles representam uma ferramenta valiosa no desenvolvimento de soluções eficientes e confiáveis em problemas de aprendizado de máquina e análise de dados.

Foram também identificados estudos que se dedicam à exploração de técnicas de aprendizado por reforço profundo Dangut *et al*., (2022) e estratégias para lidar com alta dimensionalidade (quantidade elevada de *features*) Wang, G., Chen, G. e Chu (2018). A abordagem de aprendizado por reforço profundo apresenta um potencial significativo para aprimorar a capacidade de sistemas inteligentes em aprender e tomar decisões em ambientes complexos e dinâmicos. Além disso, as técnicas desenvolvidas para lidar com a alta dimensionalidade são fundamentais para lidar com conjuntos de dados cada vez maiores e mais complexos, possibilitando a extração de informações relevantes e a identificação de padrões significativos.

As aplicações dos objetos de estudo deste contexto se concentram em resolver problemas ligados à indústria, medicina, agronegócio, análises de mercado e padrão de consumo de clientes, aviação, mercado financeiro, entre outros.

2 FUNDAMENTAÇÃO TEÓRICA

Este trabalho é caracterizado de acordo com critérios de classificação de metodologia científica uma pesquisa Exploratória de Natureza Aplicada, onde por sua vez, os conhecimentos adquiridos desempenham um papel crucial na aplicação prática e na resolução de problemas concretos da vida moderna, bem como fornecem uma compreensão mais aprofundada do problema, a fim de torná-lo explícito e facilitar a construção de hipóteses. Além do mais, esta pesquisa também possui algumas características de uma Natureza Básica, pois gera conhecimentos que podem ser utilizados em outras pesquisas. Em termos de abordagem trata-se de uma pesquisa Quantitativa, onde os resultados podem ser mensurados numericamente. Do ponto de vista dos procedimentos técnicos, é uma pesquisa combinada: Bibliográfica e Experimental, que utiliza como instrumento o desenvolvimento de uma Revisão Sistemática da Literatura (RSL) com base em protocolo criterioso e aplica os resultados obtidos para subsidiar experimentos computacionais em um objeto de estudo (MATIAS-PEREIRA, 2016).

2.1 Metodologia da Revisão Sistemática da Literatura

Uma Revisão Sistemática da Literatura é uma técnica amplamente reconhecida e utilizada para revisar e analisar trabalhos anteriores, visando obter uma visão clara e abrangente sobre um assunto específico. Essa abordagem é especialmente útil para identificar lacunas na pesquisa, abordar questões de pesquisa apresentadas em artigos existentes e avaliar os estudos disponíveis que foram publicados sobre o tema em questão (ALSOBHI *et al.*, 2023).

A realização de uma revisão sistemática é fundamental para fornecer uma visão geral atualizada do assunto e embasar a prática baseada em evidências. Para garantir uma avaliação abrangente da literatura, nossa pesquisa adotou uma abordagem sequencial baseada nos protocolos publicados por Tranfield, Denyer e Smart, (2003); Kitchenham *et al.*, (2009) e considerando apenas estudos primários relevantes que abordam a Classificação Binária em bases Extremamente Desbalanceadas. É importante ressaltar que existem muitos estudos que discutem o conceito de desbalanceamento de dados de forma geral, mas nossa revisão se concentrou exclusivamente em estudos que lidam com o desbalanceamento extremo.

Durante o processo de revisão, todos os artigos encontrados foram minuciosamente revisados e avaliados criticamente, a fim de se garantir a confiabilidade e a qualidade dos

resultados obtidos. Dessa forma, nossa Revisão proporciona uma análise aprofundada e embasada sobre Classificação Binária em bases Extremamente Desbalanceadas, contribuindo para o avanço do conhecimento nessa área e auxiliando na tomada de decisões por meio de evidências. A Figura 1 demonstra as etapas envolvidas na condução desta Revisão Sistemática da Literatura.

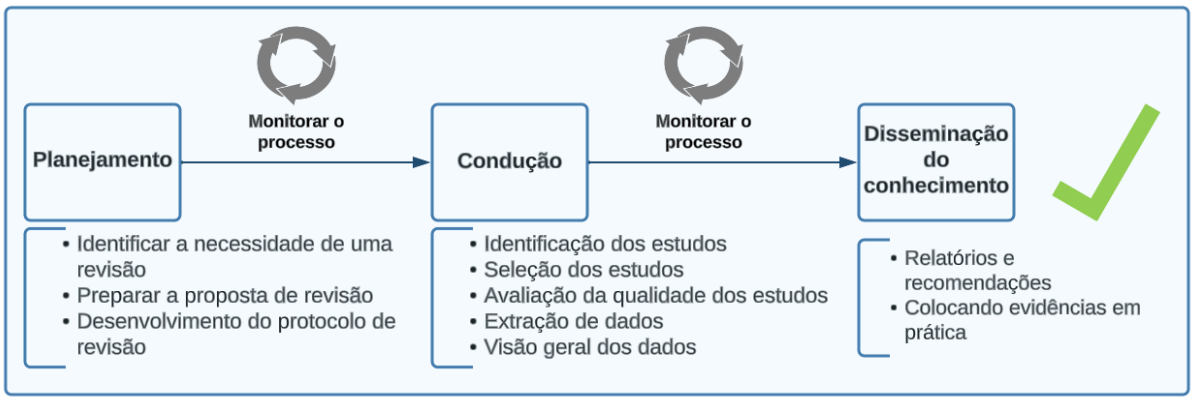
2.1.1 Estágio I - Planejamento

2.1.1.1 Fase 0: Identificação da necessidade de uma revisão

De acordo com Tranfield, Denyer e Smart, (2003), a primeira fase é verificar se já existem ou não trabalhos do tipo Revisão Sistemática da Literatura atuais e se há realmente a necessidade de se desenvolver uma. De tal modo, para o prosseguimento às próximas fases, as seguintes hipóteses serão testada a partir dos resultados obtidos:

H_0: É necessário desenvolver uma Revisão sobre o tema especificado.
H_1: Não é necessário desenvolver uma Revisão sobre o tema especificado.

Figura 1 - Etapas da Revisão Sistemática da Literatura



Fonte: Adaptado de Tranfield, Denyer e Smart (2003)

Para avaliar as hipóteses, uma busca será realizada para identificar a existência ou não de estudos com o mesmo propósito de nosso trabalho, ou seja, desenvolver uma Revisão Sistemática da Literatura sobre Classificação Binária em bases de dados Extremamente Desbalanceadas. Neste contexto, com objetivo de aumentar a abrangência da pesquisa, a busca foi realizada na plataforma online *Scopus* [2]. O critério para análise de relevância foi a verificação da quantidade de citações, a classificação do periódico, o tipo de

[2] Scopus. <https://www.scopus.com/search>.

documento, ano de publicação maior ou igual a 2019 e uma avaliação qualitativa simplificada relacionando a metodologia e os resultados obtidos pelo estudo.

Após algumas tentativas e análises, uma *String* de busca foi desenvolvida e aplicada para verificação em Novembro de 2023:

TITLE-ABS-KEY (("extreme imbalanced data") OR ("extreme unbalanced data") OR ("highest imbalanced data") OR ("highest unbalanced data")) AND TITLE-ABS-KEY ("binary classification") AND TITLE-ABS-KEY (("systematic literature review") OR ("systematic literature mapping") OR ("meta-analysis"))

Não foi encontrado nenhum artigo do tipo Revisão Sistemática da Literatura que abordasse o assunto específico. Deste modo, nós observamos que a hipótese H_0 é aceita, sendo assim existe a necessidade quanto ao desenvolvimento de uma Revisão.

Em resumo, esta Revisão Sistemática da Literatura tem como objetivo avaliar e extrair dados de:

- estudos aplicados em bases de dados Extremamente Desbalanceadas;
- estudos primários (experimentais);
- estudos sem restrição quanto a área de aplicação.

2.1.1.2 Fase 1: Preparação da proposta de revisão

Nesta fase ocorre a delimitação do escopo em relação ao tema a ser pesquisado. Para esta Revisão Sistemática da Literatura, a especificação do tema é encontrar na literatura através de estudos primários o estado da arte sobre Abordagens de Classificação Binária em bases de dados Extremamente Desbalanceadas a fim de se melhorar a precisão dos modelos. O resultado da delimitação do escopo é o desenvolvimento das Perguntas de Pesquisa (PP), exibidas no Quadro 1.

Quadro 1 - Perguntas de Pesquisa.

#	Pergunta	Justificativa
PP1	Quais as características da(s) base(s) de dados utilizadas para classificação em bases extremamente desbalanceadas?	Identificiar a proporção da classe minoritária, variáveis independentes, número de observações de treinamento e teste, entre outras características relevantes.
PP2	Qual abordagem foi utilizada para lidar com bases extremamente	Identificar as rotinas e técnicas de pré-processamento, classificadores e

	desbalanceadas?	*ensembles* frequentemente utilizados na classificação em bases extremamente desbalanceadas.
PP3	Quais as métricas de precisão avaliadas na classificação em bases extremamente desbalanceadas?	Identificar quais as métricas de precisão foram utilizadas para avaliar o desempenho da classificação em bases extremamente desbalanceadas.
PP4	Como a abordagem utilizada impacta no desempenho da classificação em bases extremamente desbalanceadas?	Analisar e/ou comparar a abordagem através das métricas de desempenho na classificação em bases extremamente desbalanceadas.

Fonte: Autor

2.1.1.3 Fase 2: Desenvolvimento do protocolo de revisão

De acordo com Kitchenham *et al.*, (2009), um protocolo de revisão especifica os métodos que serão empregados para realizar uma Revisão. Um protocolo pré-definido é necessário para reduzir a possibilidade de viés do pesquisador. Esta fase descreve como cada artigo foi identificado. Os artigos foram extraídos de uma lista de bases eletrônicas, descritas no Quadro 2, para ampliar a cobertura e abrangência das buscas. Para identificação, utilizamos o processo convencional de busca manual por meio de uma *String* adaptada para o mecanismo de pesquisa de cada base. O período de interesse será em artigos com ano de publicação maior ou igual a 2019.

Quadro 2 - Bases.

Bases de dados eletrônicas	
IEEE Xplore	https://ieeexplore.ieee.org/
SpringerLink	https://link.springer.com/
ScienceDirect	https://www.sciencedirect.com/
Web of Science	https://clarivate.com/webofsciencegroup/
Google Scholar	https://scholar.google.com.br/
Wiley Online Library	https://onlinelibrary.wiley.com/
ACM Digital Library	https://dl.acm.org/

IET Software Digital Library	https://digital-library.theiet.org/
Scopus	https://www.scopus.com/
DBLP Data Base	https://dblp.org/
PubMed	https://pubmed.ncbi.nlm.nih.gov/

Fonte: Autor

Para a execução das buscas, uma *Substring* foi desenvolvida a partir da *String* utilizada na fase de avaliação da necessidade de se desenvolver uma Revisão, onde extraímos apenas a parte relacionada a bases Extremamente Desbalanceadas:
TITLE-ABS-KEY ("extreme imbalanced data" OR "extreme unbalanced data" OR "highest imbalanced data" OR "highest unbalanced data").

2.1.2 Estágio II – Condução

2.1.2.1 Fase 3: Identificação da pesquisa

Durante o período compreendido entre 08/09/22 e 09/09/22, foram efetuadas buscas com a *string* adaptada, demonstrada na seção anterior. Os resultados obtidos foram registrados na Tabela 2, que apresenta os quantitativos de artigos encontrados por cada base.

Tabela 2 - Resultado das buscas.

Base	Quantidade de artigos
Scopus	6
IEEE Explore	1
SpringerLink	7
ScienceDirect	10
Google Scholar	60
Web of Science	3
Wiley Online Library	1
ACM Digital Library	2
IET Software Digital Library	2

DBLP Data Base	30
PubMed	1
Total	**123**

Fonte: Autor

2.1.2.2 Fase 4: Seleção dos estudos

O objetivo desta etapa é realizar a seleção dos artigos que utilizaremos para responder às Perguntas de Pesquisa descritas anteriormente no Quadro 1. Para isso, foi adotado um conjunto de Critérios de Inclusão (CI) e Critérios de Exclusão (CE) para avaliação dos artigos selecionados, que podem ser vistos no Quadro 3 e Quadro 4 respectivamente.

Após a análise do título, resumo e palavras-chave, um artigo é excluído caso atenda pelo menos um dos critérios de exclusão, ou mantido caso atenda pelo menos um dos critérios de inclusão. Em situações de dúvida, será realizada uma avaliação mais detalhada do artigo completo. Essa abordagem visa garantir que apenas os estudos mais relevantes sejam considerados, assegurando a qualidade e a confiabilidade dos resultados obtidos.

Quadro 3 - Critérios de Inclusão.

CI 1	Artigos que avaliam sistematicamente abordagens já conhecidas (técnica de pré-processamento, e/ou algoritmo de classificação, e/ou *ensemble*), aplicadas em bases extremamente desbalanceadas.
CI 2	Artigos que fornecem comparações entre abordagens já conhecidas (técnica de pré-processamento, e/ou algoritmo de classificação, e/ou *ensemble*), aplicadas em bases extremamente desbalanceadas.
CI 3	Artigos que descrevem o desenvolvimento de novas abordagens (técnica de pré-processamento, e/ou algoritmo de classificação, e/ou *ensemble*) aplicadas em bases extremamente desbalanceadas.

Fonte: Autor

Quadro 4 - Critérios de Exclusão.

CE 1	Artigos que não contemplam dados de estudos primários.
CE 2	Artigos que abordam outras metodologias para a previsão (regressão, detecção de desvios, regras de associação, padrões sequenciais, agrupamento, sumarização entre outras), ou seja, que não tenham como foco principal a CB. Também estão excluídos artigos que lidam com classificação multiclasse (CM).
CE 3	Artigos que lidam com bases de dados que não são categorizadas como extremamente desbalanceadas.

Fonte: Autor

Antes de serem submetidas aos critérios, os 123 artigos encontrados passaram por uma análise de limpeza de dados, onde ocorreu a exclusão de publicações duplicadas e daquelas nas quais não foi possível a obtenção de acesso para leitura.

Após a execução da limpeza, por meio da aplicação do conjunto de CI e CE, foram selecionados apenas 32 artigos (26%) para a próxima fase, na qual, através da avaliação da qualidade na fase subsequente, outra seleção foi realizada.

2.1.2.3 Fase 5: Avaliação da qualidade dos estudos

A avaliação da qualidade é um aspecto essencial em uma Revisão Sistemática da Literatura, uma vez que complementa o processo anterior de inclusão/exclusão de artigos com base nos Critérios de Inclusão e Exclusão. Esta avaliação tem como objetivo determinar a relevância e o rigor metodológico dos artigos selecionados. Embora não haja uma definição padrão para avaliar a qualidade de um estudo, ela geralmente consiste em uma lista de requisitos que são usados para avaliar cada artigo individualmente e classificá-los com base em um *score* (KITCHENHAM *et al.*, 2009).

É importante ressaltar que a avaliação da qualidade é fundamental para garantir a validade e a confiabilidade dos resultados obtidos em uma Revisão. Portanto, neste trabalho a avaliação foi realizada de forma criteriosa e rigorosa, a fim de garantir a qualidade desta que é a última seletiva de artigos. Nesse sentido, foi desenvolvida pelos autores uma métrica de *score* da qualidade com base em duas Questões Qualitativas (QQ):

QQ_1 - O estudo é publicado em um periódico reconhecido?

- Para esta análise foi utilizada a métrica *Percentile* do *CiteScore rank* 2021 disponibilizada pela *Scopus*. O cálculo é baseado no número de citações de documentos por um periódico ao longo de quatro anos, dividido pelo número dos mesmos tipos de documentos indexados pela *Scopus* e publicados naqueles mesmos quatro anos. Por exemplo, o *CiteScore* 2021 conta as citações recebidas em 2018-2021 por estudos publicados no mesmo período e divide isso pelo número desses documentos publicados em 2018-2021 [3].
- O *score* será aplicado de acordo com os critérios da Tabela 3 para área de estudo "Engenharias" ou outras áreas afins quando não houver classificação para a área.

Tabela 3 - *Score por percentile.*

Percentile	***Score***
>= 75th	+3
>= 50th e < 75th	+1,5
< 50th	+1
Sem avaliação	+0

Fonte: Autor

QQ_2 - As técnicas e métodos utilizados são explicadas em detalhes?

- Esta análise leva em conta se é possível reproduzir as abordagens aplicadas ao experimento com as informações documentadas no estudo (técnica de pré-processamento, e/ou algoritmo de classificação, e/ou *ensemble*, detalhamento das bases dados utilizadas (número de amostras, % da classe minoritária, *features*, entre outras), métricas de avaliação dos modelos de classificação e resultados.
- Para este *score*, os valores são baseados em três respostas, demonstradas na Tabela 4.

[3] Elsevier, How are CiteScore metrics used in Scopus? Scopus: Access and Use Support Center. <https://www.elsevier.com/products/scopus/metrics/citescore>.

Tabela 4 - *Score por detalhamento técnico/metodológico.*

Respostas	***Score***
[4]Sim	+3
[5]Parcialmente	+1,5
[6]Não	+0

Fonte: Autor

Após uma leitura mais aprofundada dos artigos com foco nas questões (QQ_1 e QQ_2), os artigos que obtiveram pontuação final (QQ_{total}) igual ou superior à metade da pontuação mais alta (QQ_{max}) foram selecionados, conforme a notação:

$$QQ_{total} = QQ_1 + QQ_2$$
$$Artigo\ selecionado,\ se\colon\ QQ_{total} \geq \frac{QQ_{max}}{2} \quad (2)$$

De tal modo, mais 10 artigos foram removidos. Permanecendo então para a sequência final deste trabalho um grupo de 22 artigos selecionados, que representam 18% do total de artigos encontrados na Fase 2.

Os níveis de qualidade estratificados dos artigos selecionados e removidos com base na pontuação obtida foram tabulados e podem ser vistos na Tabela 5.

A partir de uma análise de dispersão do *score* realizada nos 22 artigos remanescentes, foi possível constatar um valor médio do *score* (μ) de 4,9 pontos, indicando um bom nível de qualidade de acordo com os critérios estabelecidos. A variância (σ^2), por sua vez, foi relativamente baixa com valor de 1,1, indicando uma certa homogeneidade entre as pontuações obtidas.

É fundamental ressaltar a importância da utilização de critérios de qualidade para garantir a confiabilidade e relevância dos artigos científicos. No geral, os resultados indicam um bom nível de qualidade em geral com base nos critérios avaliados pela QQ_1 e QQ_2.

[4] Informações completas documentadas no estudo.
[5] Ausência ou imprecisão de no máximo um dos itens descritivos (técnica de pré-processamento, e/ou algoritmo de classificação, e/ou ensemble, bases dados, métricas de avaliação e resultados).
[6] Ausência ou imprecisão de mais de um dos itens descritivos acima.

Tabela 5 - Pontuações e níveis de qualidade dos estudos selecionados.

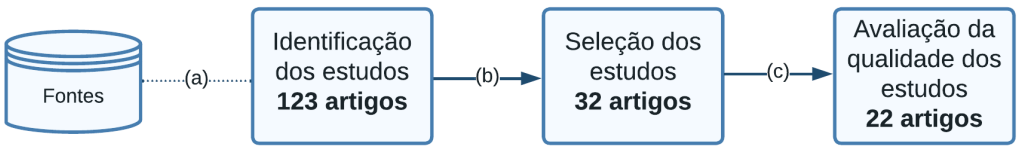
Nível de Qualidade	Quantidade de artigos	Percentual
Muito alto (5 < score <= 6)	9	28%
Alto (4 < score <= 5)	7	22%
Médio (3 <= score <= 4)	6	19%
*Baixo (0 < score < 3)	10	31%
Total	32	100%

* Artigos removidos.

Fonte: Autor

Ao término desta fase, obtivemos os artigos selecionados para a extração de dados. A Figura 2 descreve os resultados deste processo de seleção e avaliação dos estudos através da aplicação de filtros em cada fase. Onde (a) corresponde a aplicação da String de Busca desenvolvida na Fase 2: Desenvolvimento do protocolo de revisão; (b) os critérios de inclusão e exclusão e (c) a aplicação da avaliação da qualidade dos estudos com base na QQ_1 e QQ_2.

Figura 2 - Resultados do processo de seleção de estudos



Fonte: Autor

2.1.2.4 Fase 6: Extração dos dados

Nesta etapa, nosso objetivo é obter o conhecimento necessário para responder às Perguntas da Pesquisa apresentadas no Quadro 1. Para alcançar esse objetivo, realizamos uma leitura minuciosa dos estudos selecionados, seguindo o formulário de extração de

dados apresentado no Quadro 5. Essa abordagem nos permitiu extrair informações relevantes e fundamentais para a nossa análise e fornecer respostas embasadas e precisas às Perguntas de Pesquisa. A extração de dados é um processo fundamental para consolidar e sintetizar o conhecimento existente, de forma objetiva e evitando o desvio do escopo da pesquisa.

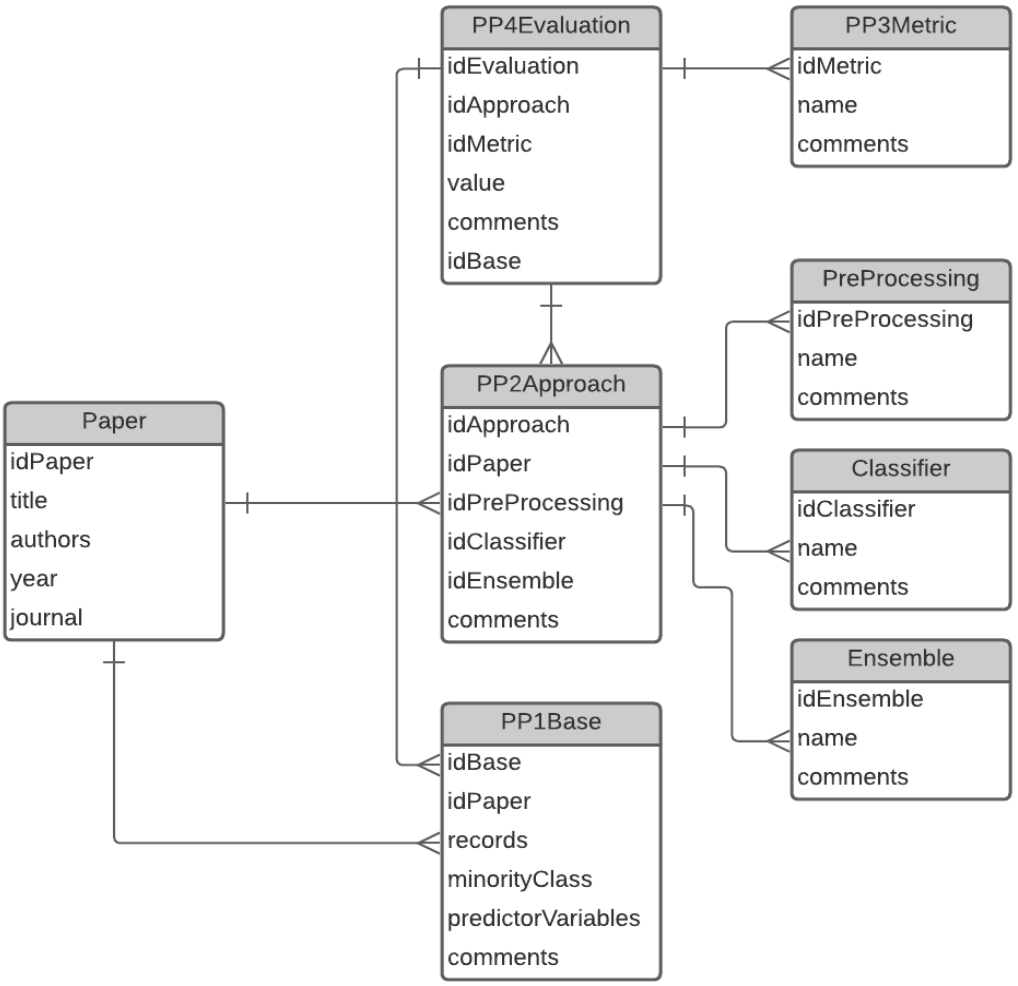
Quadro 5 - Formulário de Extração de Dados.

ID do Estudo: **Título:** **Autore(s):** **Ano:** **Journal:** **PP1 - Quais as características da(s) base(s) de dados utilizadas para classificação em bases extremamente desbalanceadas?** ● Detalhamento da(s) base(s) de dados. **PP2 - Qual abordagem foi utilizada para lidar com bases extremamente desbalanceadas?** ● Identificar pré-processamento, classificadores e modelos combinados (ensembles). ● Descrever a abordagem utilizada. **PP3 - Quais as métricas de precisão avaliadas na classificação em bases extremamente desbalanceadas?** ● Métricas de precisão utilizadas. **PP4 - Como a abordagem utilizada impacta no desempenho da classificação em bases extremamente desbalanceadas?** ● Valores das métricas obtidos com a abordagem utilizada.

Fonte: Autor

Para auxiliar na análise de dados, as respostas obtidas através do Formulário de Extração de Dados foram armazenadas em um banco de dados relacional *PostgreSQL* v12.16, cuja a modelagem é apresentada na Figura 3. Por meio desta estrutura é possível organizar as informações e obter consultas e cruzamentos de dados para subsidiar os resultados e discussões através da linguagem de consultas SQL. Esta base de dados é parte da etapa de Disseminação do Conhecimento prevista no protocolo da RSL.

Figura 3 - Modelagem da base de dados.



Fonte: Autor

2.1.2.5 Fase 7: Síntese dos dados

Esta fase sumariza os resultados das Perguntas de Pesquisa que foram extraídos pela aplicação do Formulário de Extração de Dados. Como o objetivo é consolidar os dados extraídos, os subtítulos subsequentes apresentam a síntese da extração de dados, pois respondem através da análise de dados às Perguntas de Pesquisa. Além do mais, fornecem resultados analíticos, *insights* e apontamentos sobre os principais achados desta Revisão Sistemática da Literatura. A ferramenta utilizada foi o Formulário de Extração de Dados, apresentado no Quadro 5.

PP1 - Quais as características da(s) base(s) de dados utilizadas para CB em bases extremamente desbalanceadas?

Ao todo, foram identificadas 52 bases de dados utilizadas nos experimentos dos artigos selecionados. As características observadas nas bases de dados para nosso estudo incluem: i. percentual da classe minoritária; ii. quantidade de variáveis preditoras (*features*) e iii. número de amostras. A Tabela 6 apresenta um resumo geral destes aspectos.

Tabela 6 - Resumo geral das bases de dados

i. Percentual da classe minoritária (menor - maior)	0,001% - 1%
ii. Número de variáveis preditoras (menor - maior)	1 - 74
iii. Número de Amostras (menor - maior)	81 - 4.976.391

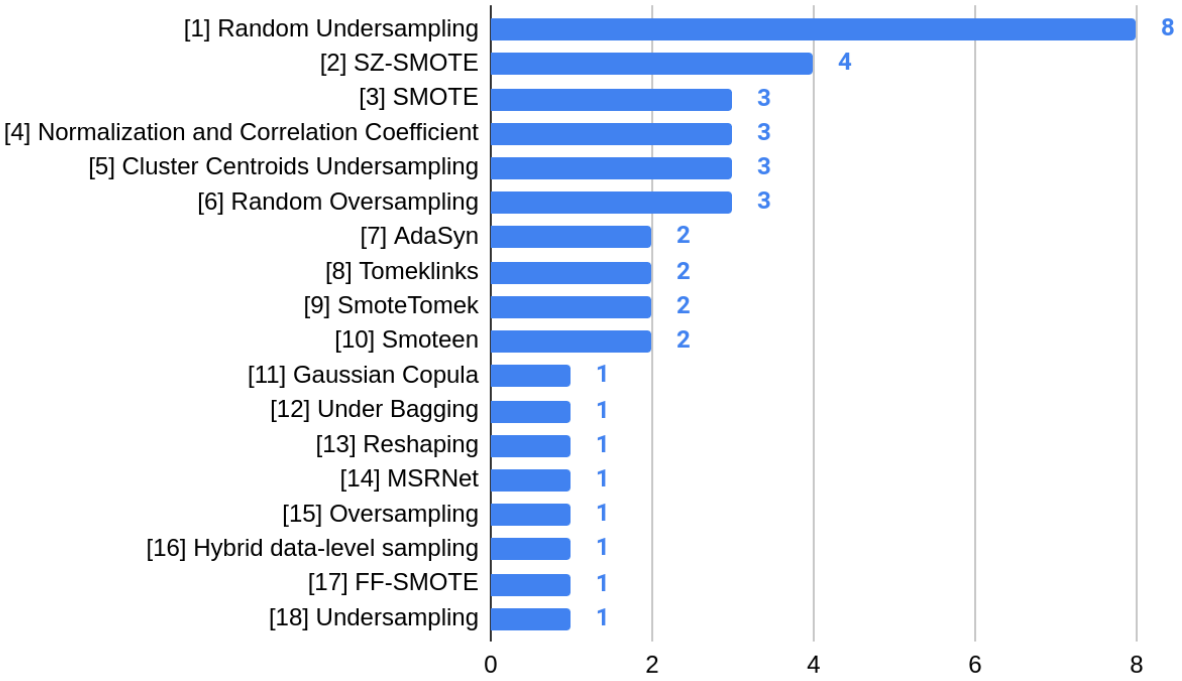
Fonte: Autor

Observamos primeiramente que todas as bases de dados encontram-se muito próximas ao limiar da definição Extremamente Desbalanceadas, pois possuem Percentual da classe minoritária >= 1%. Adicionalmente, os itens Número de Variáveis Preditoras e Número de Amostras apresentam uma larga amplitude, o que para é muito importante para garantir que as análises abrangem vários cenários e desafios diferentes em aplicações de Classificação Binária em bases Extremamente Desbalanceadas.

PP2 - Qual abordagem foi utilizada para lidar com bases extremamente desbalanceadas?

As abordagens extraídas dos artigos foram empregadas para lidar com as bases Extremamente Desbalanceadas. Para facilitar o entendimento e as análises iniciais, agrupamos os achados em: i. Técnicas de Pré-processamento; ii. Algoritmos de Classificação; e iii. *Ensembles,* apresentando informações quantitativas sobre o uso das abordagens que obtiveram melhor desempenho nos experimentos reportados em cada um dos artigos selecionados. Os gráficos exibidos nas Figuras 4, 5 e 6 sumarizam as abordagens mais utilizadas em formato de *ranking*.

Figura 4 - Técnicas de Pré-processamento



[1] (RAGHUWANSHI e SHUKLA, 2019); (PEREIRA, P. J. *et al.*, 2021); (SINGH, RANJAN e TIWARI, 2021); (RAGHUWANSHI e SHUKLA, 2019b); (YU *et al.*, 2019).

[2] (WANG, D. *et al.*, 2021).

[3] (PEREIRA, P. J. *et al.*, 2021); (SINGH, RANJAN e TIWARI, 2021).

[4] (TRISANTO *et al.*, 2021).

[5, 6, 7, 8, 9, 10] (SINGH, RANJAN e TIWARI, 2021).

[11] (PEREIRA, P. J. *et al.*, 2021).

[12] (TRAN, T., TRAN, L. e MAI, 2019).

[13] (VELANDIA-CARDENAS, VIDAL e POZO, 2021).

[14] (XIA *et al.*, 2021).

[15] (SIKDER *et al.*, 2022).

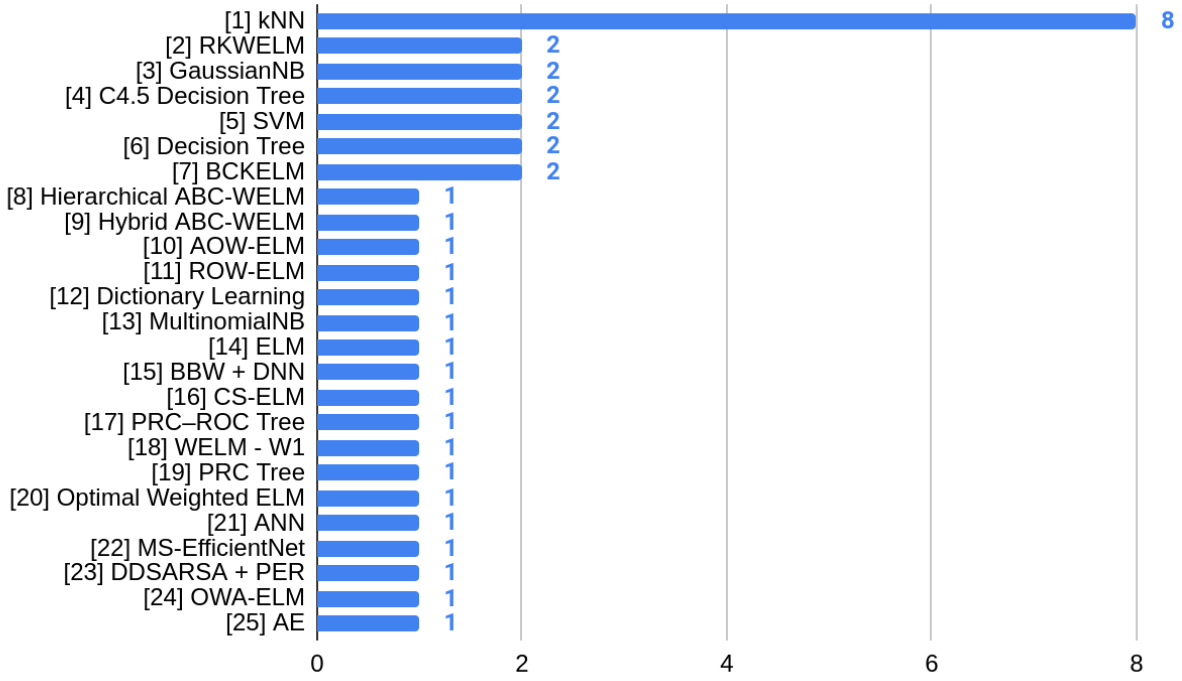
[16] (KAUR e GOSAIN, 2020).

[17] (KAUR e GOSAIN, 2019).

[18] (LAQUEUR *et al.*, 2022).

Fonte: Autor

Figura 5 - Algoritmos de classificação



[1] (SINGH, RANJAN e TIWARI, 2021); (TRISANTO *et al.*, 2021); (WANG, D. *et al.*, 2021).

[2] (RAGHUWANSHI e SHUKLA, 2019).

[3, 13] (SINGH, RANJAN e TIWARI, 2021).

[4] (KAUR e GOSAIN, 2019); (KAUR e GOSAIN, 2020);

[5] (VELANDIA-CARDENAS, VIDAL e POZO, 2021); (SINGH, RANJAN e TIWARI, 2021).

[6] (SINGH, RANJAN e TIWARI, 2021); (WANG, D. *et al.*, 2021).

[7] (RAGHUWANSHI e SHUKLA, 2019b).

[8] (RAGHUWANSHI e SHUKLA, 2019).

[9] (TANG e CHEN, L., 2019)

[10, 11, 14] (YU *et al.*, 2019).

[12] (CHEN, Y. e SHAYILAN, 2022).

[15] (HU, J. *et al.*, 2021).

[16, 18, 24] (ZHANG, X. e QIN, 2022).

[17, 19] (MIAO e ZHU, W., 2021).

[20] (LU, C. *et al.*, 2019).

[21] (WANG, D. *et al.*, 2021).

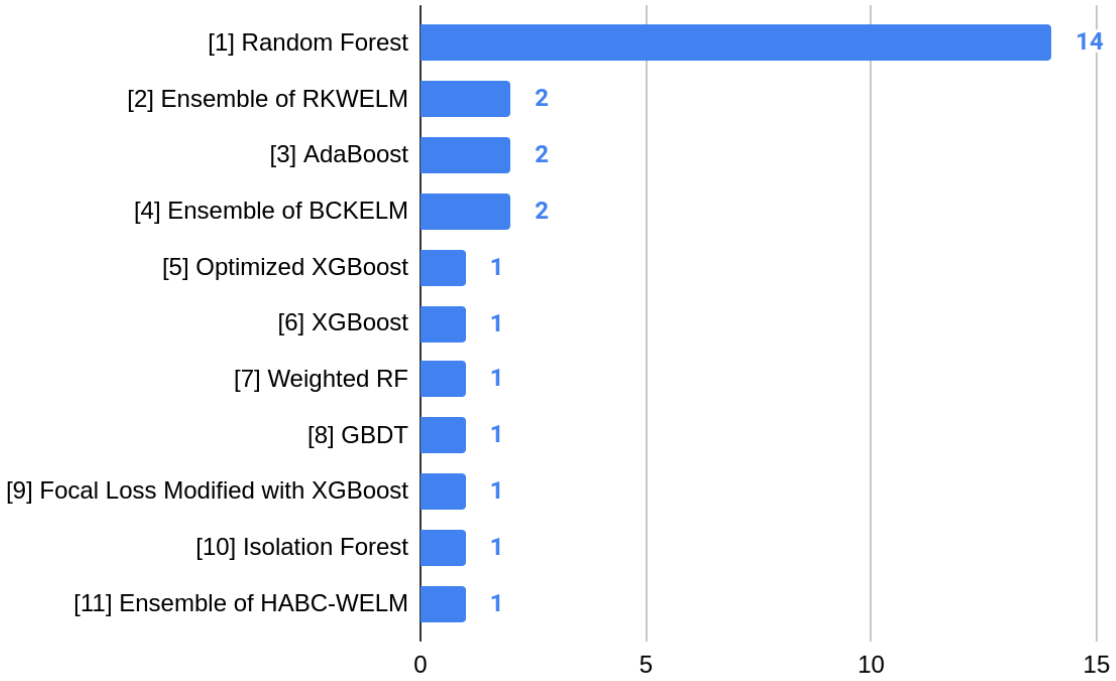
[22] (XIA *et al.*, 2021).

[23] (DANGUT *et al.*, 2022).

[25] (FONTES *et al.*, 2022).

Fonte: Autor

Figura 6 - *Ensembles*



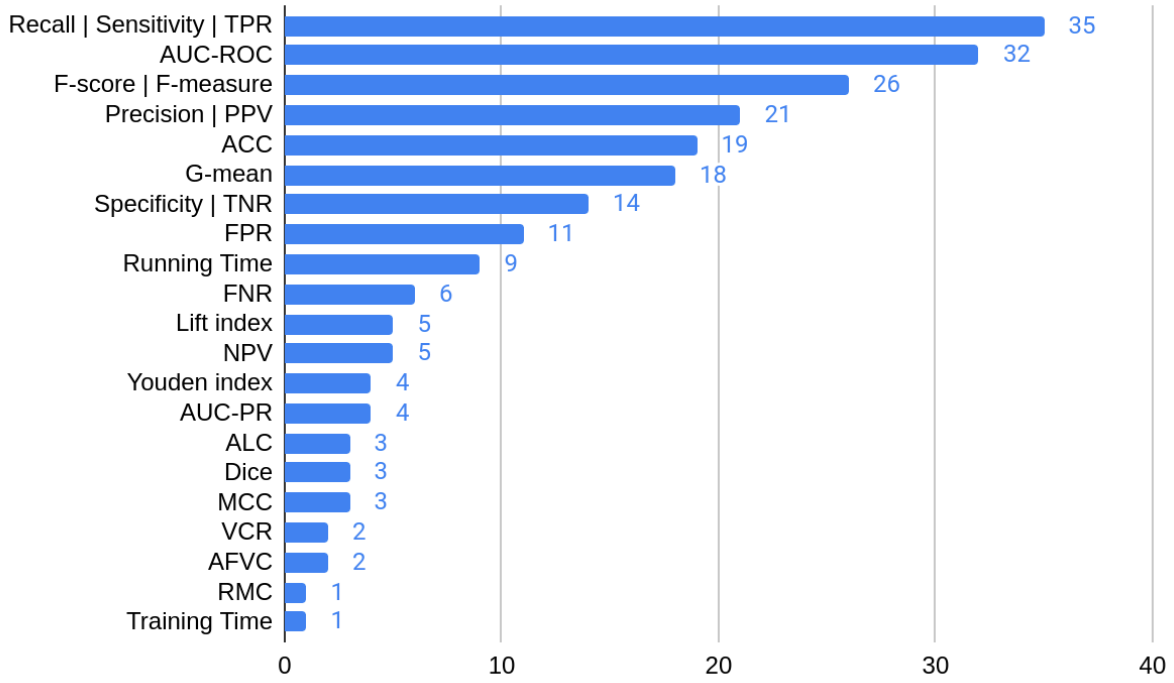
[1] (LAQUEUR *et al.*, 2022); (TRAN, T., TRAN, L. e MAI, 2019); (PEREIRA, P. J. *et al.*, 2021); (SIKDER *et al.*, 2022); (FONTES *et al.*, 2022); (SINGH, RANJAN e TIWARI, 2021); (WANG, D. *et al.*, 2021).
[2, 11] (RAGHUWANSHI e SHUKLA, 2019).
[3, 6] (SINGH, RANJAN e TIWARI, 2021).
[4] (RAGHUWANSHI e SHUKLA, 2019b).
[5, 9] (TRISANTO *et al.*, 2021).
[7] (MIAO e ZHU, W., 2021).
[8] (WANG, D. *et al.*, 2021).
[10] (FONTES *et al.*, 2022).
Fonte: Autor

PP3 - Quais as métricas de precisão avaliadas na classificação em bases extremamente desbalanceadas?

Nesta resposta, apresentaremos de forma quantitativa e sumarizada as métricas de precisão utilizadas na avaliação das abordagens. Para isso, consideramos a quantidade de estudos que as utilizaram e exibimos na Figura 7. É importante ressaltar que a avaliação das abordagens é fundamental para garantir a eficácia e a qualidade dos resultados obtidos. De tal modo, é necessário utilizar métricas de precisão que permitam mensurar de forma objetiva o desempenho das soluções propostas. Todavia é importante destacar que a

escolha das métricas adequadas depende do problema em questão e das características dos dados utilizados.

Figura 7 - Métricas



Fonte: Autor

PP4 - Como a abordagem utilizada impacta no desempenho da classificação em bases extremamente desbalanceadas?

As análises de dados são essenciais para nos ajudar a compreender como as diferentes abordagens afetam o desempenho obtido pelas métricas em relação ao Percentual da Classe Minoritária. Embora todas as bases de dados possam ser categorizadas como Extremamente Desbalanceadas, analisamos na Tabela 6 uma considerável amplitude no Percentual (P) variando de 0,001% até 1%. De tal modo, para refinar as análises, os resultados foram agrupados em quatro faixas proporcionais, que chamaremos de Cenários de Desbalanceamento Extremo (CDE) demonstrados na Tabela 7.

É importante salientar que em todos os artigos selecionados as abordagens foram testadas experimentalmente e os resultados foram extraídos e armazenados para a obtenção das análises e discussões apresentadas.

Tabela 7 - Cenários de Desbalanceamento Extremo

CDE	Faixa (P)	Artigos	Bases	$\overline{(P)}$
(I)	P < 0,25	(XIA et al., 2021); (DANGUT *et al*., 2022); (LAQUEUR *et al*., 2022); (TRAN, T., TRAN, L. e MAI, 2019); (PEREIRA, P. J. *et al*., 2021); (VELANDIA-CARDENAS, VIDAL e POZO, 2021); (SIKDER *et al.*, 2022); (FONTES *et al*., 2022); (HU, J. *et al*., 2021); (CHEN, Y. e SHAYILAN, 2022); (SINGH, RANJAN e TIWARI, 2021); (TRISANTO *et al*., 2021).	31	0,09
(II)	0,25 >= P < 0,5	(TRISANTO *et al*., 2021).	1	0.39
(III)	0,5 >= P < 0,75	(RAGHUWANSHI e SHUKLA, 2019); (HU, J. *et al.*, 2021).	3	0.50
(IV)	P >= 0,75	(RAGHUWANSHI e SHUKLA, 2019); (TRAN, T., TRAN, L. e MAI, 2019); (MIAO e ZHU, W., 2021); (RAGHUWANSHI e SHUKLA, 2019b).; (KAUR e GOSAIN, 2019); (WANG, D. *et al.*, 2021); (KAUR e GOSAIN, 2020); (LU, C. *et al*., 2019); (YU *et al.*, 2019); (ZHANG, X. e QIN, 2022); (TANG; CHEN, L., 2019)	17	0.87

Fonte: Autor

A Tabela 8 exibe as abordagens que alcançaram o melhor desempenho em cada cenário e faz referência ao respectivo artigo, considerando a abordagem que obteve o maior valor dentre as seis primeiras métricas apresentadas na Figura 7: *Recall (TPR), AUC-ROC*

(AR), F-score (F-S), Precision (PPV), ACC e G-Mean (G), quem em termos quantitativos, foram utilizadas em mais da metade dos estudos selecionados.

Em *Machine Learning,* a métrica *Recall* também referida como *True Positive Rate (TPR)*, Taxa de Verdadeiros Positivos, ou *Sensitivity* é usada para medir a porcentagem de positivos reais que são identificados corretamente (WANG, H.; ZHENG, 2013). Sendo sua fórmula:

$$Recall = \frac{TP}{TP + FN}, \text{ onde:} \quad (3)$$

TP = Número de verdadeiros positivos (identificados corretamente).

FN = Número de falsos negativos (rejeitados incorretamente).

A métrica *Area under the Receiver Operating Characteristic Curve (AUC-ROC)*, Área sob a Curva de Característica de Operação do Receptor, mede toda a área bidimensional abaixo da curva ROC inteira, como no cálculo integral de (0,0) a (1,1), conforme exemplificado na Figura 8. No entanto, a curva ROC é um gráfico que mostra o desempenho de um modelo de classificação em todos os limites de classificação, representada por dois parâmetros: *TPR* e *FPR*. Sendo o FPR, *False Positive Rate,* ou Taxa de Falsos Positivos, calculado pelo fórmula:

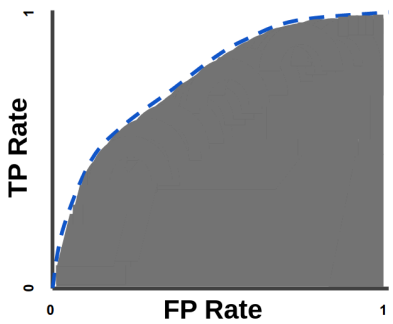
$$FPR = \frac{FP}{FP + TN}, \text{ onde:} \quad (4)$$

FP = Número de falsos positivos (identificados incorretamente).

TN = Número de verdadeiros negativos (identificados corretamente).

AUC é um algoritmo eficiente e baseado em classificação que oferece uma medida agregada de desempenho em todos os limites de classificação possíveis. *AUC* varia no valor de 0 a 1. Um modelo com previsões 100% incorretas tem uma *AUC* de 0,0, e uma com previsões 100% corretas tem uma *AUC* de 1,0 (GOOGLE FOR DEVELOPERS, 2022).

Figura 8 - AUC (Área sob a curva ROC).



Fonte: Google for Developers (2022)

F-Score ou *F1-Score* é uma maneira de combinar as métricas *Precision* e *Recall* do modelo, e é definida como a média harmônica delas, sendo uma forma de avaliar o desempenho de um modelo de classificação binária (MURPHY, 2012). Sendo sua fórmula:

$$F1 = \frac{2 * P * R}{R + P} \text{ , onde:} \quad (5)$$

P = *Precision* (Valor Preditivo Positivo).

R = *Recall* (Taxa de Verdadeiros Positivos).

É possível também ajustar a *F-Score* para dar mais peso à *P* do que à R, ou vice-versa. Sendo as *F-Scores* ajustadas mais comuns são a *F0.5-Score* e a *F2-Score*, além da *F1-Score* padrão.

A métrica *Precision*, *Positive Predictive Value* (*PPV*), Valor Preditivo Positivo, ainda segundo Murphy (2012) mede qual fração de nossas detecções é realmente positiva. Ou seja, a proporção de instâncias classificadas como positivas que são realmente positivas em relação ao total de instâncias classificadas como positivas, incluindo as verdadeiras positivas e os falsos positivos. Sendo sua fórmula:

$$Precision = \frac{TP}{TP + FP} \text{ , onde:} \quad (6)$$

TP = Número de verdadeiros positivos (identificados corretamente).

FP = Número de falsos positivos (identificados incorretamente).

De acordo com Velandia-Cardenas, Vidal e Pozo (2021) a métrica *Accuracy* (*ACC*), Acurácia, mede o número de previsões corretas feitas pelo modelo sobre o número total de observações pela fórmula:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$

Accuracy é uma métrica simples e intuitiva, mas nem sempre é a mais adequada para avaliar de forma isolada um modelo de classificação, especialmente em situações de desequilíbrio de classes. O ideal é utilizá-la em conjunto com outras métricas, como *Precision*, *Recall* e *F1-Score*, para obter uma avaliação mais abrangente do desempenho do modelo de classificação.

A métrica *G-Mean*, Média Geométrica, é especialmente útil em conjuntos de dados desbalanceados. Ele fornece uma medida agregada do desempenho do modelo, considerando tanto a capacidade de detectar corretamente a classe minoritária (*Recall* ou *Sensitivity*) quanto a capacidade de evitar classificar erroneamente instâncias da classe

majoritária como pertencentes à classe minoritária (*True Negativa Rate, TNR* ou *Specificity*) (RAGHUWANSHI e SHUKLA, 2019a). Para encontrar a *Specificity*, aplicamos a fórmula:

$$Specificity = \frac{TN}{TN + FP} \quad (8)$$

De tal modo para calcular a *G-Mean,* temos:

$$G - Mean = \sqrt{Sensitivity * Specificity} \quad (9)$$

Uma pontuação de G-mean mais alta indica um melhor desempenho do modelo, onde valores mais próximos de 1 são ideais

Também são apresentados detalhes sobre a base de dados em que a abordagem de melhor desempenho foi aplicada: Percentual da Classe Minoritária (P), Quantidade de Amostras (A) e Número de *Features* (*F*). As análises dos quatro cenários indicam que as abordagens combinadas apresentam um desempenho superior em seus experimentos. O fato pode ser explicado pela complementaridade entre as diferentes abordagens.

i. Análises do CDE I

O CDE I é o cenário com as condições mais desafiadoras. De acordo com os resultados encontrados nas melhores abordagens, o estudo de Velandia-Cardenas, Vidal e Pozo (2021) propõe uma metodologia de detecção de falhas na operação e manutenção de fazendas eólicas aplicadas a dados reais do *Supervisory Control And Data Acquisition* (SCADA). Três estratégias de pré-processamento foram testadas: *Principal Component Analysis* (PCA) para modelagem e redução de dados; *Random Oversampling*; e a técnica de *Reshaping* de dados para aumento da quantidade de informação por amostra. Também foi avaliado um *Time-Split* para evitar corromper a estrutura temporal do conjunto de dados (quando os dados são dependentes do tempo) e prevenir dispersão de dados durante o treinamento dos algoritmos de Aprendizagem de Máquina. A combinação dessas técnicas de pré-processamento de dados leva a um excelente desempenho dos algoritmos de classificação. Dentre os testes, a abordagem que utilizou a técnica de pré-processamento *Reshaping* com o algoritmo de classificação *Support Vector Machine* (SVM) e sem *Oversampling* obteve os melhores resultados. Além disso, se as observações de falhas forem separadas por intervalos de tempo mais longos, o *Reshaping* pode melhorar o problema de dados desequilibrados. Os testes realizados identificaram a janela de tempo de

1 hora como a melhor para esta aplicação. O estudo de Singh, Ranjan e Tiwari (2021) investiga várias estratégias de tratamento de Desequilíbrio Extremo, visando aprimorar os sistemas de detecção de transações fraudulentas de cartão de crédito. Experimentos extensivos em uma base de dados pública de transações de cartão de crédito[7] avaliam várias abordagens e os resultados sugerem que a combinação de técnicas de pré-processamento baseadas em *Oversampling*, como por exemplo *Adaptive Synthetic Sampling* (*AdaSyn*) podem melhorar modelos de *ensemble*, como o *Random Forest* (RF) e outros. Em comparação, o estudo Velandia-Cardenas, Vidal e Pozo (2021) do mesmo CDE, lidou com um desafio maior em relação à quantidade de amostras (A), aplicando os experimentos em uma base aproximadamente dez vezes menor que o apresentado pelo estudo de Singh, Ranjan e Tiwari (2021) e com Percentual (P) 0,07 pontos menor. Por outro lado, o estudo Singh, Ranjan e Tiwari (2021) teve que lidar com o número de *features* (F) 50% maior comparado ao estudo de Velandia-Cardenas, Vidal e Pozo (2021), que também é um fator que influencia significativamente a qualidade dos modelos. Levando em conta os aspectos mencionados e os resultados das métricas, podemos deduzir que ambas as abordagens são equivalentes e podem apresentar boa performance para outras aplicações em bases de dados que se enquadrem no CDE I, todavia é importante considerar também as especificidades e objetivos de cada tipo de aplicação, bem como outras características das bases de dados.

ii. Análises do CDE II

O estudo de Trisanto *et al.* (2021) foi o único que se enquadrou na faixa percentual do CDE II. Os autores propõem uma abordagem, um *ensemble* denominado *Modified Focal Loss method for Imbalanced XGBoost (MFL XGBoost)*, adicionando um parâmetro de desequilíbrio (φ), cujo o valor é obtido da função *W-CEL* (*Weighted – Cross Entropy Los*s) para melhorar a capacidade da função *Focal Loss,* baseando-se no conceito de *Weighted Binary Cross-Entropy*. Os autores utilizaram a mesma base de dados de detecção de transações fraudulentas de cartão de crédito do estudo de Singh, Ranjan e Tiwari (2021), entretanto com os dados originais a abordagem proposta teve desempenho inferior. Porém, os autores derivaram outros dois cenários a partir dos dados originais, reduzindo os registros com rótulos não-fraude (classe majoritária). Sendo um cenário com P = 4,68% (fora da área de interesse do nosso trabalho) e outro com P = 0,39%. Neste último, a abordagem citada anteriormente obteve os melhores resultados nas métricas. Não foi utilizado como

[7] Detecção de fraude de cartão de crédito. Disponível em: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

abordagem de pré-processamento, técnicas de *sampling* ou detecção de *outlier*. Foram aplicados a Normalização (da *feature* "*amount*") e o Coeficiente de Correlação para encontrar as *features* mais significativas para o modelo. A abordagem proposta pelos autores demonstrou resultados satisfatórios na redução de vieses e na confiabilidade dos modelos, contudo os próprios autores destacam ainda que mais pesquisas são necessárias para melhorar a performance em dados com desequilíbrio mais extremo, como visto nos estudos selecionados para o CDE I.

Tabela 8 - Abordagens com melhor desempenho em CDE

		Base			**Métricas**					
CDE	**Abordagens**	**P**	**F**	**A**	**TPR**	**AR**	**F-S**	**PPV**	**ACC**	**G**
(I)	*Reshaping* + *SVM* (VELANDIA-CARDENAS; VIDAL e POZO, 2021).	0,10	20	29.488	**1**	NC	**1**	**1**	**1**	NC
	AdaSyn + *RF* (SINGH; RANJAN e TIWARI, 2021).	0,17	30	284.807	**1**	NC	**1**	**1**	**1**	NC
(II)	*NCC* + *FLM XGBoost* (TRISANTO *et al*., 2021)	0,39	30	125.492	**0,72**	NC	NC	**0,97**	**1**	NC
(III)	*BBW* + *DNN* (HU, J. *et al.*, 2021)	0,5	1	402	NC	NC	**0,99**	**0,99**	NC	NC
	RUS + *RKWELM* + *Ensemble of RKWELM* (RAGHUWANSHI; SHUKLA, 2019a)	0,5	2	1.005	NC	**1**	NC	NC	NC	**0,99**
(IV)	*SZ-SMOTE* + *RF* (WANG, D. *et al*., 2021)	0,93	32	750	**1**	**1**	**0,99**	**0,99**	**0,99**	NC

*NC = Não consta avaliação a para a métrica. Fonte: Autor

iii. Análises do CDE III

No CDE III os dois estudos que se enquadraram, Hu, J. *et al.* (2021); Raghuwanshi e Shukla (2019a) obtiveram resultados bem próximos em seus experimentos, embora em métricas diferentes. No estudo de Hu, J. *et al.* (2021) a base de dados selecionada é denominada *BonnEEG,* sendo utilizada para a detecção de epilepsia baseada em EEG (Eletroencefalografia), um método de monitoramento eletrofisiológico que é utilizado para registrar a atividade elétrica do cérebro. Com registros obtidos por reamostragem e com apenas uma *feature*, dados em *single-channel*. Um novo *framework* é proposto com objetivo de adaptar um algoritmo de classificação baseado em uma Rede Neural Profunda genérica para ser melhor treinada a partir de conjuntos de dados extremamente desequilibrados com poucas amostras minoritárias. Com auxílio do *Batch Balance Wrapper* (BBW), duas camadas de rede extras são adicionadas ao início de uma Rede Neural Profunda. As camadas evitam o *overfitting* de amostras minoritárias e melhoram a expressividade da distribuição. Além disso, o *Batch Balance* (BB), um algoritmo de amostragem baseado em classes, é proposto para garantir que as amostras em cada lote sejam sempre balanceadas durante o processo de aprendizado. Conjuntos de dados desequilibrados representam um desafio para Redes Neurais Profundas, mas o BBW pode ajudar efetivamente no seu treinamento. Comparativamente às abordagens existentes, verificou-se que o BBW alcançou uma melhor performance na classificação dos dados utilizados nos experimentos realizados neste estudo. Para a execução dos experimentos e avaliação foi utilizado um método para conduzir uma validação cruzada (*leave-one-out*), adaptada com objetivo de reduzir o tempo de execução. Adicionalmente, as Redes Neurais Profundas encapsuladas pela técnica do BBW demonstraram ser 16,39 vezes mais rápidas do que as não encapsuladas nos testes realizados no âmbito deste estudo. A aplicação do BBW não requer pré-processamento adicional dos dados ou ajuste extra dos hiperparâmetros, operações estas que normalmente consumiram um tempo maior.

No estudo de Raghuwanshi e Shukla (2019a), os resultados obtidos na faixa de interesse do CDE III foram aplicados em uma base de dados sintética, com duas *features* e geradas usando uma distribuição *Gaussiana* G(0.7, 0.5). Os autores abordam um *ensemble* de *Reduced Kernelized Weighted Extreme Learning Machine* (RKWELM) e como técnica de pré-processamento utilizam *Random Undersampling* (RUS). Os KWELM reduzidos do classificador base são treinados de forma sequencial. Em cada iteração, as amostras classificadas corretamente pertencentes à classe majoritária são substituídas pelas outras classes majoritárias para criar um novo subconjunto de *kernel* balanceado. Os resultados do conjunto criado de classificadores são combinados usando a Votação Majoritária (MV) e a Votação Suave (SV). Este trabalho usa WELM com função *kernel* Gaussiana para mapear

os dados de entrada para o espaço de características. Os melhores resultados obtidos variando o parâmetro de regularização foram no intervalo {2^{-18}, 2^{-16}, …2^{48}, 2^{50}}. O parâmetro de largura do *kernel*, é pesquisado no intervalo {2^{-18}, 2^{-16}, …2^{18}, 2^{20}} para encontrar o resultado otimizado. De forma similar ao estudo de Hu, J. *et al.* (2021) também foi utilizada validação cruzada (*fivefold*) para cada *dataset*, de tal modo os valores dos resultados possuem uma média e desvio padrão. Uma observação significativa, em comparação a outros estudos, é que ambos os estudos selecionados para este CDE III, embora apresentem abordagens robustas para lidar com o Desbalanceamento Extremo, poderiam ter explorado os resultados em um conjunto mais amplo de métricas de avaliação.

iv. Análises do CDE IV

No estudo de Wang, D. *et al.* (2021), os autores propõem uma técnica de pré-processamento estendida do *Synthetic Minority Over-sampling Technique* (SMOTE), o *Safe Zone (*SZ*) SMOTE*. Eles comparam com outras técnicas de pré-processamento combinadas a diferentes algoritmos de classificação isolados e *ensembles*. O método *SZ-SMOTE* gera amostras sintéticas com um mecanismo de concentração na área Hiperesférica ao redor de cada instância minoritária selecionada. Sendo este procedimento resumido em três objetivos:

i. Para definir uma Zona Segura flexível em torno de cada instância da classe minoritária selecionada, de forma que as instâncias minoritárias geradas dentro dessa área não sejam ruidosas. A Zona Segura é uma hiperesfera com a amostra minoritária como centro e a distância da amostra majoritária mais próxima à amostra minoritária como raio; ii) O raio da zona segura é definido como base para a geração da distribuição de amostras minoritárias; iii) Para aumentar a variedade de amostras geradas, a densidade da classe minoritária é expandida. Este processo visa aumentar a representatividade da classe minoritária, contribuindo para a melhoria do desempenho de algoritmos de classificação em bases de dados Extremamente Desbalanceados. Os resultados apresentados foram obtidos pela aplicação dos experimentos em uma base de dados para previsão de terremotos. Dados de monitoramento de 10 de junho de 2017 a 30 de junho de 2019 capturados por um equipamento do tipo *Acoustic and Electromagnetic Testing All* (AETA) na província de *Sichuan* - China. Mesmo que uma área esteja numa zona propensa a terremotos, a proporção de dias com terremotos por ano ainda é uma minoria. De acordo com o que foi apresentado pelos autores, os resultados do experimento mostram que a qualidade da previsão de terremotos usando SZ-SMOTE supera significativamente o de usar outras

técnicas do *oversampling*. Além disso, a abordagem combinada que obteve os melhores resultados foi o *Ensemble Random Forest* (RF) .

2.1.3 Estágio III - Disseminação do Conhecimento

Lidar com o desbalanceamento extremo apresenta desafios significativos em sua aplicação. Os resultados desta Revisão Sistemática da Literatura apontam alguns direcionamentos que podem auxiliar pesquisadores e profissionais da área, com destaque para a escolha inicial das abordagens (*kick off*) ajustadas de acordo com as características das bases de dados.

Analisando os artigos com melhor performance geral, apresentados na Tabela 8, observamos que ao utilizar mais de uma técnica, é possível explorar as vantagens de cada uma delas e minimizar suas limitações, aumentando a robustez da abordagem. Dessa forma, é possível obter resultados mais precisos e confiáveis no contexto da Classificação Binária em bases de dados Extremamente Desbalanceadas. Em complemento, os artigos da Tabela 8 e suas respectivas abordagens apresentam qualidade considerável em se tratando dos critérios (*score*): Relevância (QQ_1) e Replicabilidade (QQ_2). Dentre os seis artigos, três foram classificados com nível de qualidade Muito Alto (RAGHUWANSHI; SHUKLA, 2019a); (VELANDIA-CARDENAS; VIDAL e POZO, 2021); (HU, J. *et al.*, 2021), dois como Alto (SINGH; RANJAN e TIWARI, 2021); (TRISANTO *et al*., 2021) e um como Médio (WANG, D. *et al*., 2021).

De forma geral, os principais achados desta pesquisa são apresentados nos subtópicos a seguir.

2.1.3.1 Fase1: Relatórios e Recomendações

i. Tipos de abordagens mais utilizadas

Analisando o quantitativo total dos 22 artigos selecionados, sem levar em conta as melhores performances no geral, as técnicas de Pré-processamento foram aplicadas em 40 abordagens, os algoritmos de Classificação em 38 abordagens e os *Ensembles* em 27 abordagens.

Em vários artigos foram testadas abordagens combinando técnicas de pré-processamento e algoritmos de classificação, o que justifica o quantitativo bem próximo entre as abordagens. Já os *ensembles* em valor quantitativo menor, representam

abordagens um pouco mais complexas em termos de aplicação e custo computacional, de tal modo, são utilizados em cenários experimentais mais bem definidos e elaborados. Por outro lado, ao analisarmos os tipos empregados pelas abordagens de melhor performance geral da Tabela 8, observamos que o padrão da análise quantitativa se mantém em partes, com destaque ao uso de *ensembles*. Analisando isoladamente, cinco abordagens utilizaram técnicas de Pré-processamento, quatro utilizaram *Ensembles* e duas algoritmos de Classificação. A grande maioria das abordagens de melhor performance utilizaram combinação entre tipos, com destaque significativo para os uso de técnicas de Pré-processamento com *Ensembles*, abordagem presente em quatro dos seis estudos.

ii. Técnicas/métodos mais utilizados

Iniciando pela análise dos quantitativos totais mais expressivos, vemos a técnica de Pré-processamento *Random Undersampling* (RUS) ser empregada em 8 abordagens, o algoritmo de classificação *K-Nearest Neighbor* (kNN) também em 8 abordagens e o *ensemble Random Forest* (RF) em 14 abordagens. Ambas as técnicas são amplamente difundidas na literatura e largamente utilizadas por profissionais nas mais diversas áreas de estudo em problemas de Aprendizagem de Máquina. Ao analisarmos as abordagens de melhor performance no geral, destacamos o *Ensemble* RF, presente em dois (SINGH; RANJAN e TIWARI, 2021); (WANG, D. *et al.*, 2021) dos seis artigos nos respectivos CDE (I) e (IV). Além do mais, nestes dois cenários, RF performou bem utilizado em conjunto com técnicas derivadas de *Oversampling* (*AdaSyn* e *SZ-SMOTE*).

iii. Principais achados catalogados

As informações obtidas por meio do Formulário de Extração de Dados foram armazenadas em um banco de dados relacional, cujo a modelagem é ilustrada na Figura 3. Através dessa estrutura as informações puderam ser melhor organizadas. Consultas e análises cruzadas puderam ser realizadas para fundamentar os resultados e discussões. Além disso, essa estrutura possibilitará a descoberta de novas ideias e poderá ser atualizada com novos resultados no futuro.

2.1.3.2 Fase 2: Colocando Evidências em Prática

Dada a importância de se evidenciar e consolidar os principais achados desta RSL, o detalhamento das ações que envolvem a execução dos experimentos, bem como de seu objeto de estudo serão apresentados a partir do próximo capítulo promovendo a conexão lógica entre teoria e prática desta pesquisa.

3 MATERIAIS E MÉTODOS

Nesta seção será apresentada a metodologia utilizada para o planejamento e execução dos experimentos, bem como de seu objeto de estudo. As aplicações experimentais complementam os achados da RSL e são parte da fase Colocando Evidências em Prática, conforme o protocolo abordado na seção anterior.

Observamos significativo destaque em relação às abordagens exploradas pela fundamentação teórica, o uso de técnicas de pré-processamento em conjunto com *ensembles*. Mais especificamente técnicas derivadas de *Oversampling* combinadas com o *ensemble Random Forest* (RF) obtiveram as melhores performances nos cenários de desbalanceamento extremo. Para validar na prática por meio da experimentação, nesta seção será apresentada a metodologia para o desenvolvimento do objeto de estudo, composto por bases de dados sintéticas geradas dentro das condições de desbalanceamento extremo. Em complemento, a seção subsequente (Cap. 4) apresenta a avaliação da performance dos modelos de classificação obtidos através da aplicação das abordagens destacadas pela RSL, bem como analisa e discute os resultados obtidos.

3.1 Metodologia para geração de bases sintéticas

Dados do mundo real, abrangendo uma variedade de tipos, como texto, imagens, vídeos e áudio, frequentemente apresentam desafios devido a questões como incompletude, preocupações com a confidencialidade, restrições regulatórias e dificuldades de aquisição. Além dessas limitações, a necessidade de geração controlada de dados sintéticos levou pesquisadores a explorar diversas metodologias que permitem a avaliação acelerada de algoritmos e métodos de aprendizado de máquina. Esses desafios impulsionaram o avanço das técnicas de geração de dados sintéticos, cada vez mais reconhecidas por sua capacidade de apoiar o progresso na pesquisa em aprendizado de máquina e inteligência artificial (HAO et al., 2024).

Além disso, dados sintéticos ganharam destaque como uma alternativa segura a dados reais em contextos sensíveis à privacidade. De acordo com Murtaza et al. (2023), conjuntos de dados sintéticos são cada vez mais empregados em inúmeros projetos de pesquisa devido à sua capacidade de apoiar análises e compartilhamento de informações sem comprometer informações pessoais ou sensíveis. Essa vantagem de privacidade torna os dados sintéticos particularmente valiosos em cenários onde o acesso a dados reais é restrito por requisitos de confidencialidade. Como resultado, os dados sintéticos não apenas garantem a continuidade da pesquisa em domínios sensíveis, mas também contribuem para

o cumprimento da legislação de privacidade de dados, ampliando assim sua aplicabilidade em pesquisas de IA e outras disciplinas com uso intensivo de dados.

De tal modo, este trabalho apresenta uma abordagem baseada em Planejamento de Experimentos (DoE) para a geração de conjuntos de dados sintéticos extremamente desbalanceados, que serão utilizados como objeto de estudos. Ao oferecer um padrão estruturado para a geração de dados sintéticos, a abordagem proposta permite avaliações controladas e aceleradas do desempenho algorítmico sem a necessidade de aquisição inicial de dados reais.

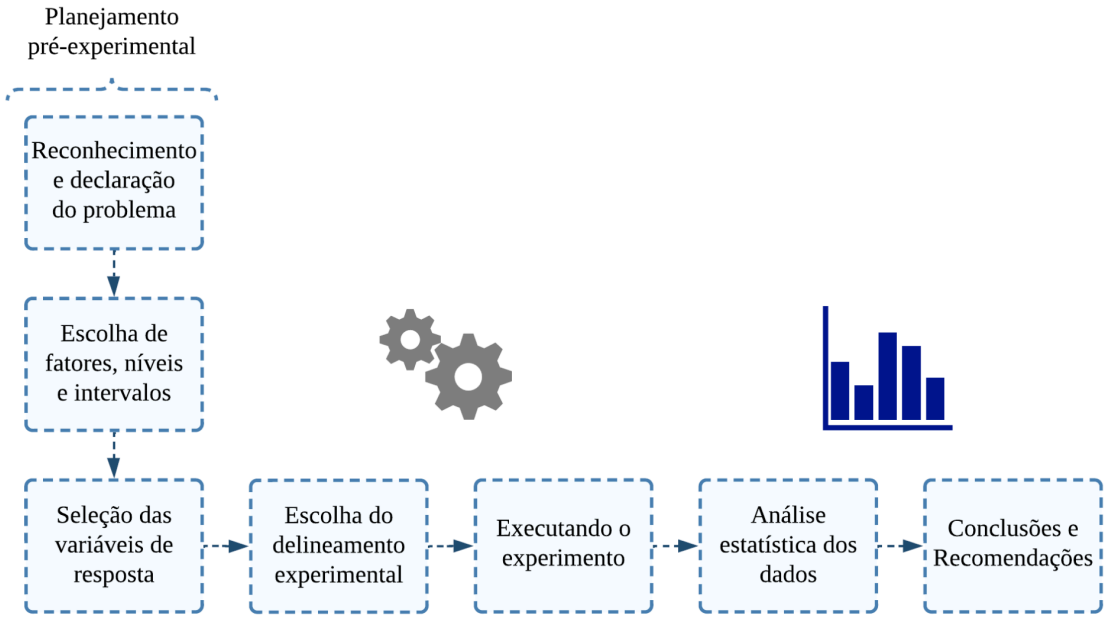
As bases sintéticas extremamente desbalanceadas serão utilizadas no treinamento e teste de modelos de classificação binária, sendo posteriormente avaliados pela utilização de seis indicadores de desempenho: *TPR, PPV, AUC-ROC, Accuracy, F1-Score e G-MEAN*, permitindo uma análise multidimensional do comportamento e desempenho do classificador. São as mesmas métricas de performance obtidas e utilizadas na RSL.

3.1.1 Abordagem baseada em Planejamento de Experimentos (DoE)

O Planejamento de Experimentos ou *Design of Experiments* (DoE), também denominado planejamento estatístico de experimentos, refere-se ao processo de planejamento do experimento de modo que os dados apropriados sejam coletados e analisados por métodos estatísticos, resultando em conclusões válidas e objetivas. A abordagem estatística é essencial para extrair conclusões significativas dos dados, especialmente quando estes estão sujeitos a erros experimentais, tornando os métodos estatísticos a única abordagem objetiva de análise. Existem dois aspectos em qualquer problema experimental: o design do experimento e a análise estatística dos dados. Estes dois tópicos estão intimamente relacionados, uma vez que o método de análise depende diretamente do design utilizado (MONTGOMERY, 2012).

De acordo com Dean *et al*. (2017), o propósito de um experimento pode ser caracterizado por exploratório e confirmatório, de tal modo, a filosofia da análise depende do propósito do experimento. A abordagem escolhida para condução do planejamento estatístico de experimentos apresentado neste artigo, segue as diretrizes de Montgomery (2012), conforme ilustrado na Figura 9 e será detalhada nos subtópicos seguintes.

Figura 9 - Diretrizes para projetar um experimento.



Fonte: Montgomery (2012).

3.1.1.1 Reconhecimento e declaração do problema

Esta abordagem tem por objetivo a geração de dados sintéticos classificados como Extremamente Desbalanceadas que não dependam de dados reais preexistentes. Através do *DoE*, uma experimentação controlada pode ser conduzida em conjunto com uma avaliação estatística baseada em seis métricas apropriadas para modelos de Classificação Binária em conjuntos de dados com desequilíbrio extremo. Além do mais, as bases geradas podem auxiliar pesquisas por novos algoritmos de classificação de forma mais acelerada, sem a necessidade inicial de se coletar dados reais de problemas de classes extremamente desbalanceadas.

3.1.1.2 Escolha de fatores, níveis e intervalos

De acordo com Montgomery (2012) os fatores de projeto são os fatores realmente selecionados para o estudo, os quais pode-se desejar variar no experimento. Uma vez selecionados, deve-se escolher as faixas de valores nas quais esses fatores serão variados e os níveis específicos nos quais as execuções serão realizadas.

Para este experimento, os fatores foram selecionados dentre os aspectos descritivos mais comuns observados em conjuntos de dados extremamente desbalanceados em problemas de classificação binária de diversas áreas. O Quadro 5 descreve cada um dos fatores.

Quadro 5 - Fatores selecionados.

Fator	Descrição
A. *Features*	Conjunto de variáveis independentes ou preditoras (dimensionalidade) que caracterizam e descrevem um dado problema. Cada uma dessas informações, incluídas na representação de um objeto de estudo, é denominada uma *feature* (GOODFELLOW et al., 2016).
B. *Sampling*	Quantidade de amostras geradas. Para este experimento o valor definido será multiplicado pelo valor de *Features*, sendo o produto a quantidade total de amostras da base sintética. Por exemplo, para *Features* = 2 e *Sampling* = 1000, teremos uma base de dados com 2000 amostras (2 * 1000).
C. *Extreme Unbalance Percentage*	Valor do percentual da Classe Minoritária. Apresenta o nível de desbalanceamento da base. Consiste na dificuldade associada aos dados que possuem uma grande maioria de registros pertencentes a uma única classe (majoritária) em detrimento a outra (minoritária) (DATAR; GARG, 2019).
D. *Type of function for Response (Y)*	Apresenta o tipo de função utilizada para gerar cada y em função das variáveis independentes (**x**), de acordo com a representação $y = f(x)$.
E. *Threshold*	Limiar que arredonda a variável de resposta (y). Em problemas de classificação binária Y assume um de dois valores, por exemplo (0,1) (JORDAN; MITCHELL, 2015). Para este experimento 1 representa a classe minoritária, de tal modo, os resultados de y obtidos através de um função são arredondados para 1 conforme o valor do *Threshold* selecionado.
F. *Error*	Valor que representa o termo de erro da função. Corresponde a variabilidade que o modelo não consegue explicar, ou seja, a parte da variabilidade nos dados que não é atribuída às variáveis independentes. O erro (ϵ), que de acordo com Montgomery (2012) é denominado termo de erro aleatório, para este experimento possui média (μ) = 0 e os níveis variam o percentual do desvio padrão (σ) dos valores de y.

Fonte: Autor.

As faixas de valores dos fatores *Features, Sampling e Extreme Unbalance Percentage* se encontram foram definidas após análise de 52 bases de dados classificadas como Extremamente Desbalanceadas, referenciadas no Capítulo 2 desta Tese. Contudo, a

escolha dos níveis específicos em que os fatores foram submetidos neste experimento foram baseadas também em critérios observacionais (levando-se em conta experiências e práticas comuns em aplicações que envolvem *Machine Learning*), respeitando-se as faixas limites analisadas na literatura.

Os demais fatores, relacionados à lógica utilizada para obtenção da variável de resposta binária, foram selecionados em conformidade com conceitos e práticas relativas a funções matemáticas e estatísticas básicas. O conjunto de fatores e níveis é apresentado em detalhes na Tabela 9.

Tabela 9 - Detalhamento de Fatores e Níveis.

Fator	**Nível -**	**Nível +**	**Descrição**
A	2	10	Para o **nível -** a base terá 2 *Features* e para o **nível +** 10 *Features***,** que representam a quantidade de variáveis independentes (**x**) da base de dados. Os níveis foram escolhidos com valores quantitativos equilibrados. De acordo com Abdulrauf Sharifai e Zainol (2020) a classificação de um conjunto de dados desequilibrado torna-se mais complexa na medida em que o número de *Features* se torna muito grande.
B	1000	1500	Valor que, multiplicado ao número de *Features*, retorna a quantidade de amostras da base sintética. Para o **nível -** o valor do multiplicador é [1000 x fator A] e para o **nível +** é [1500 x fator A]. De tal modo, a menor base possui 2000 amostras e a maior 15000 amostras. Um número limitado de amostras torna difícil a classificação (ABDULRAUF SHARIFAI; ZAINOL, 2020), bem como, um número muito alto significa em alto custo computacional, o que desencadeia muito tempo para processamento dos modelos. Para este experimento os valores do vetor (**x)** são gerados aleatoriamente por uma distribuição uniforme com intervalo entre -2 e 2. Mais detalhes na etapa Executando o Experimento.
C	0,5%	1%	Para o **nível -** 0,5% e para o **nível +** 1% de desbalanceamento. Representa a quantidade de amostras da classe minoritária. De acordo com o Google for Developers (2025), 1% é o limite que faz fronteira entre a categoria de Desbalanceamento Extremo e Moderado.
D	Linear Simples	Com interação	O **nível -** aplica uma Funções Lineares Simples para obtenção a variável dependente y, conforme notação: $y = \beta_0 + \beta_1 x_1 \ldots + \beta_i x_i$, onde i = número de *Features*. No entanto, o **nível +** aplica Funções Lineares com Interação, incluindo interações de 2ª e 3ª ordem, sendo a última apenas para bases com 10 *Features* (fator **A +**). Em modelos do mundo real, de acordo com Montgomery (2012) um número pequeno de variáveis são consideradas

			importantes e relevantes, bem como poucas interações podem ser consideradas significativas.
E	0,5	0,8	Para o **nível -**, todas as respostas de y >= 0,5 são convertidas para a classe 1 e para **nível +** todas as respostas de y >= 0,8 são convertidas para a classe 1. Em ambos os casos, as demais respostas de y são convertidas para a classe 0 (majoritária).
F	20%	80%	Para o **nível -** aplica-se 20% ao valor de (σ) e para o **nível +** aplica-se 80% ao valor de (σ). Sendo o valor do termo de erro aleatório (ϵ) uma distribuição normal com $\mu = 0$ e desvio padrão percentual em relação ao valor do desvio padrão dos valores de y. Por exemplo, para o desvio padrão de $y = 1.67$ e se o **nível -** for selecionado, teremos $(1.67 * 0.2)$, logo: (ϵ) será um valor aleatório retornado por uma distribuição normal de $\mu = 0$ e $\sigma = 0.334$ Após, o valor de y é recalculado com a adição do termo (ϵ) à função: $$y = \beta_0 + \beta_1 x_1 \ldots + \beta_i x_i + \epsilon \quad (10)$$

Fonte: Autor.

3.1.1.3 Seleção das variáveis de resposta

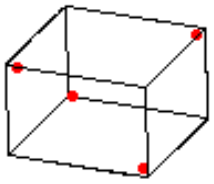
Cada experimento planejado, apresentado na Tabela 10, tem como resultado uma base de dados sintética com as características em conformidade aos fatores e níveis apresentados nas Tabelas 8 e 9. A partir da geração das bases sintéticas por meio de um *script* escrito em linguagem R, cada uma delas é submetida a um processo de Classificação Binária, através de um *script* em *Python*, que retorna um grupo de métricas de avaliação relativos à qualidade do modelo (definidas com base nos resultados da RSL): ***TPR, PPV, AUC-ROC, Accuracy, F1-Score e G-MEAN***. Tais métricas, juntamente com o tempo gasto para a geração da base - *Elapsed Time* - (variável que tem o objetivo de avaliar o custo computacional envolvido no processo), formam as variáveis de resposta deste experimento. Ainda de acordo com Montgomery (2012), muitas das vezes, a média ou o desvio padrão (ou ambos) da característica medida será a variável de resposta, bem como múltiplas respostas podem ser necessárias para avaliar o experimento adequadamente. Os detalhes dos *scripts* são demonstrados no tópico Executando o experimento.

3.1.1.4 Escolha do delineamento experimental

A seleção do design experimental envolve a consideração do tamanho da amostra (número de réplicas), a escolha de uma ordem de execução apropriada para os ensaios experimentais e a determinação da necessidade de blocos, controlar a variabilidade em fatores não controláveis ou outras restrições de randomização (MONTGOMERY, 2012). A condução deste experimento faz uso de uma abordagem Fatorial Fracionada com dois níveis para cada fator.

De acordo com a documentação da Minitab ([s.d.]) um experimento fatorial é um tipo de experimento planejado que possibilita a observação dos efeitos de múltiplos fatores sobre uma resposta. Ao conduzir o experimento, a variação simultânea dos níveis de todos os fatores, em vez de alterá-los individualmente, permite o estudo das interações entre esses fatores. Os experimentos do tipo Fatorial Fracionados, como demonstrado na Figura 10, utilizam uma fração (metade, um quarto, etc.) de todas as combinações possíveis de configurações dos fatores, com um número reduzido de execuções em comparação ao delineamento fatorial completo de 2^k. O fatorial completo deste experimento seria composto por 64 execuções ou 2^6, contudo o fatorial fracionado $\frac{1}{2}$ foi selecionado, sendo então 32 execuções realizadas.

Figura 10 - Exemplo Fatorial Fracionado $\frac{1}{2}$



Se o número de fatores for cinco ou mais, o ideal é executar um delineamento fatorial fracionado com vistas a reduzir o número de execuções, mantendo a capacidade de modelar todas as interações de dois fatores (MINITAB, [s.d.]). Ademais, o grau de resolução em um delineamento fatorial fracionado refere-se à capacidade do delineamento de distinguir entre efeitos principais e interações de diferentes ordens. É uma medida da qualidade do delineamento em termos de confusão ou aliasing de efeitos. Nosso experimento foi delineado com grau de resolução IV, adequado para identificar e estimar efeitos principais e, em certa medida, interações de dois fatores, com alguma ambiguidade em interações mais complexas.

Para auxiliar na fase de design experimental, o *software Minitab*® 21 com seu pacote de *DoE* foi utilizado. Após a definição dos fatores e níveis e seleção do fatorial fracionado, a Tabela 10 exibe o planejamento experimental com as respectivas ordens de execução, onde cada execução retornou uma base de dados sintética.

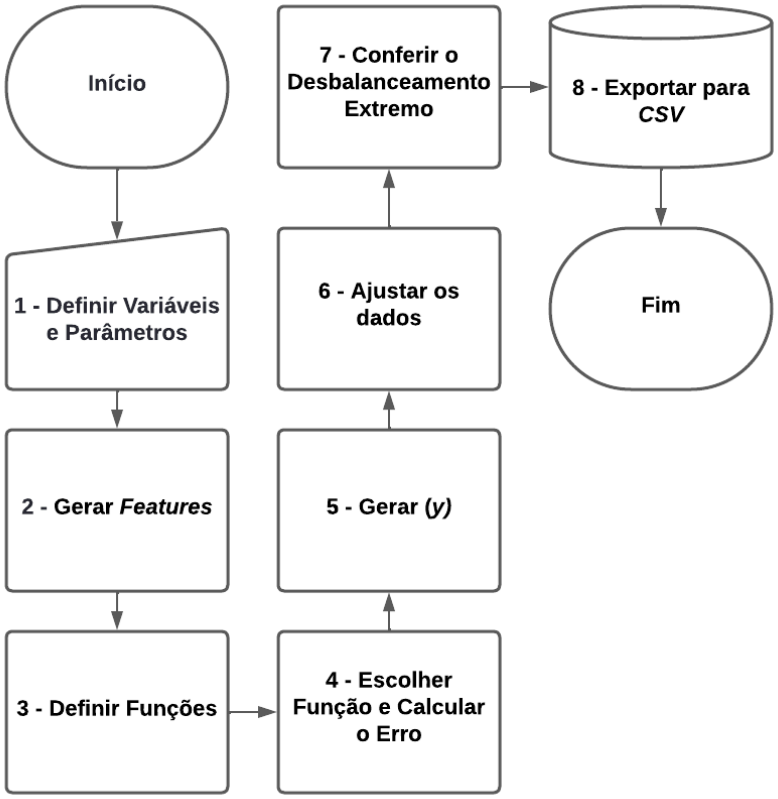
Tabela 10 - Planejamento Experimental.

Order	***#***	***Features***	***Sampling***	***Extreme Unbalance Percentage***	***Type of function for Response (Y)***	***Threshold***	***Error***
1	17	2	1000	0,5%	Linear Simples	0,8	0,8
2	15	2	1000	1%	Com interações	0,5	0,8
3	4	10	1000	0,5%	Linear Simples	0,5	0,2
4	7	2	1000	1%	Linear Simples	0,5	0,2
5	10	10	1000	0,5%	Com interações	0,5	0,2
6	8	10	1000	1%	Linear Simples	0,5	0,8
7	32	10	1000	1%	Com interações	0,8	0,8
8	6	10	1000	1%	Linear Simples	0,5	0,2
9	27	2	1000	0,5%	Com interações	0,8	0,8
10	14	10	1000	1%	Com interações	0,5	0,8
11	12	10	1000	0,5%	Com interações	0,5	0,8
12	30	10	1000	1%	Com interações	0,8	0,2
13	3	2	1000	0,5%	Linear Simples	0,5	0,8
14	25	2	1000	0,5%	Com interações	0,8	0,2
15	24	10	1000	1%	Linear Simples	0,8	0,2
16	2	10	1000	0,5%	Linear Simples	0,5	0,8
17	19	2	1000	0,5%	Linear Simples	0,8	0,2
18	16	10	1000	1%	Com interações	0,5	0,2
19	31	2	1000	1%	Com interações	0,8	0,2
20	13	2	1000	1%	Com interações	0,5	0,2
21	22	10	1000	1%	Linear Simples	0,8	0,8
22	29	2	1000	1%	Com interações	0,8	0,8
23	21	2	1000	1%	Linear Simples	0,8	0,2
24	23	2	1000	1%	Linear Simples	0,8	0,8
25	9	2	1000	0,5%	Com interações	0,5	0,8
26	1	2	1000	0,5%	Linear Simples	0,5	0,2
27	11	2	1000	0,5%	Com interações	0,5	0,2
28	18	10	1000	0,5%	Linear Simples	0,8	0,2
29	26	10	1000	0,5%	Com interações	0,8	0,8
30	28	10	1000	0,5%	Com interações	0,8	0,2
31	20	10	1000	0,5%	Linear Simples	0,8	0,8
32	5	2	1000	1%	Linear Simples	0,5	0,8

3.1.1.5 Executando o experimento

A execução dos experimentos ocorreu em duas etapas: **i.** Geração das bases sintéticas e **ii**. Aplicação de Classificação Binária nas bases geradas. Na primeira etapa, cada execução do experimento delineado gerou uma base de dados sintética com as características estabelecidas em cada um dos seis fatores. Um *script* (Anexo A e C) em linguagem R foi desenvolvido para automatizar o processo de geração das bases e cada uma das suas partes (*steps*) estão representadas na Figura 11.

Figura 11 - Fluxograma do algoritmo de geração de bases sintéticas



---

1. **Definir Variáveis e Parâmetros**: os parâmetros permitem a criação de múltiplos cenários experimentais, cada um com uma combinação única de níveis para as variáveis seguindo a ordem de execução da Tabela 10. O objetivo é observar como diferentes configurações de fatores influenciam a performance e a robustez dos modelos de *machine learning* desenvolvidos com esses dados sintéticos.

---

2. **Gerar Features:** neste *step*, são geradas as features (variáveis independentes) que compõem as amostras do conjunto de dados sintéticos $[x_1, \ldots x_n]$**.** O número total de

amostras é calculado multiplicando o número de features (Fator A) pelo multiplicador de amostras (Fator B). Este cálculo ajusta o tamanho do conjunto de dados de acordo com a dimensionalidade das features e a escala definida pelo multiplicador. A matriz de *features* (**x**) é gerada através da função *runif*, que cria uma sequência de números aleatórios uniformemente distribuídos entre -2 e 2.

---

3. **Definir Funções:** aqui são definidas as funções responsáveis por calcular a variável dependente (y) a partir das variáveis independentes (x). Essas funções são criadas para capturar diferentes cenários, com e sem interação entre as variáveis, e para conjuntos de dados com diferentes quantidades de *features* (2 ou 10).

---

4. **Escolher Função e Calcular o Erro**: neste *step,* são selecionadas as funções adequadas para gerar a variável dependente (y) com base no número de *features* (Fator A) e na presença de interações (Fator D). Além disso, é calculado o desvio padrão (*effect*) da variável y para ser utilizado na adição de ruído às funções geradoras sendo após e ajustado pelo percentual do erro escolhido (Fator F).

---

5. **Gerar (y)**: a variável dependente (y) é gerada aplicando as funções lineares previamente definidas às variáveis independentes (x). A aplicação das funções lineares (com ou sem interações) às variáveis independentes permite a geração da variável dependente (y) de acordo com as especificações experimentais. A inclusão de um termo de erro controlado assegura que a variabilidade nos dados sintéticos seja consistente com os parâmetros de erro especificados.

---

6. **Ajustar (y)**: neste *step*, os dados gerados são ajustados para introduzir desbalanceamento extremo na variável de resposta (y). Este processo envolve: **i.** o arredondamento dos valores de (y), onde valores maiores ou iguais ao *Threshold* (Fator E) são arredondados para 1, caso contrário para 0; **ii.** cálculo do desbalanceamento desejado com base no total de amostras e na taxa de *Extreme Unbalance Percentage* (Fator C); **iii.** junção dos valores de (x) e (y); **iv.** remoção de observações da classe dominante (1) para atingir o nível de desbalanceamento desejado; **v.** ajuste final para manter o número total de amostras, onde cópias das observações da classe majoritária (0) são feitas para preencher as observações removidas, garantindo que o número total de amostras permaneça o mesmo.

---

7. **Conferir o Desbalanceamento Extremo**: verifica-se se a base de dados sintética gerada atingiu o nível de desbalanceamento extremo especificado. Isso é feito

calculando a proporção da classe minoritária (1) e comparando-a com o valor de *Extreme Unbalance Percentage* (Fator C) selecionado para o experimento. A verificação do desbalanceamento extremo é um passo crucial para assegurar que a base de dados sintética gerada cumpre o principal requisito experimental.

---

8. **Exportar para CSV:** neste *step* final a base de dados sintética gerada, contendo desbalanceamento extremo conforme os parâmetros especificados, é exportada para um arquivo *CSV*. Os dados são armazenados[8] de forma persistente, o que possibilita sua utilização posterior em análises e experimentos subsequentes com abordagens de *Machine Learning*.

O código desenvolvido para geração de bases sintéticas está dentro de uma função *system.time,* que permite quantificar o tempo necessário para a execução completa do *script*. Medir o tempo de execução de um bloco de código é uma abordagem útil para avaliar a eficiência computacional de um *script.* Para este experimento, o tempo gasto em cada execução, que retorna um arquivo CSV com uma base de dados sintética, é armazenado na variável de resposta *Elapsed Time* descrita no Quadro 2. O *hardware* para execução desta etapa foi um processador *Intel® Core™* i5-5200U CPU @ 2.20GHz com 4 núcleos com 8 GB de memória *RAM* em sistema operacional *Ubuntu* 20.04.6 LTS de 64-bits. Para edição do código em R foi utilizada a *IDE RStudio* 2023.12.1+402.

Na segunda etapa da execução experimental, as bases sintéticas geradas foram submetidas à abordagens apropriadas para Classificação Binária em bases extremamente desbalanceadas. Conforme observado na Fundamentação Teórica, o *ensemble Random Forest* combinado com técnicas de pré-processamento baseadas em *Oversampling*, apresentaram resultados significativos em cenários de desbalanceamento extremo, com destaque aos trabalhos de (SINGH et al., 2021) e (WANG et al., 2021).

De acordo com Bistroń e Piotrowski (2022)*, Random Forest (RF)* baseia-se no uso de múltiplas árvores de decisão, aumentando a precisão do modelo em relação ao uso de árvores de decisão individuais. Cada árvore individual em uma floresta aleatória retorna uma previsão de pertencimento à classe selecionada, e a classe com mais votos torna-se a previsão final do modelo. Este método é amplamente utilizado devido à sua eficácia em lidar com problemas de classificação e regressão em conjunto com grandes conjuntos de dados.

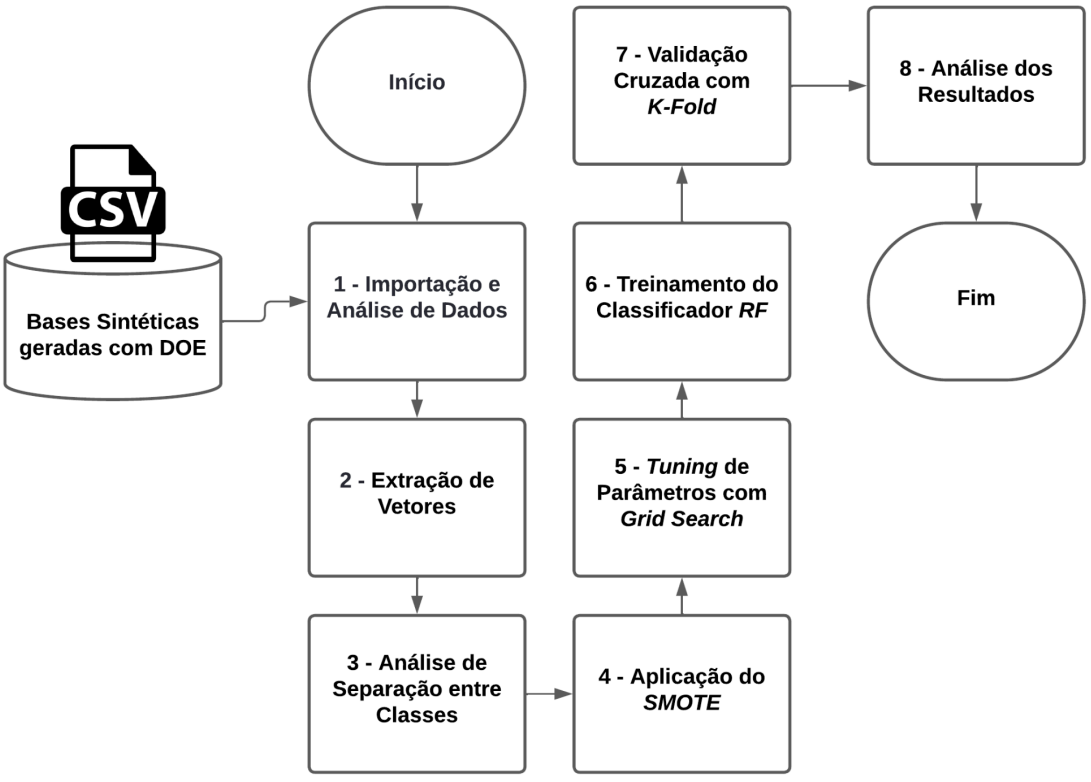
O processo de *Oversampling* duplica aleatoriamente pontos de dados minoritários para aumentar sua contagem, o que pode muitas vezes levar ao *overfitting* do modelo. No

[8] As bases sintéticas estão disponíveis em: <https://drive.google.com/drive/folders/1NbU0HYyUVah83UD8gry5lEEEvevkzeYz?usp=sharing>

entanto, o *SMOTE*, uma técnica mais avançada de *Oversampling* que será utilizada neste experimento em conjunto ao *Random Forest*, cria novas amostras da classe minoritária para equilibrar a distribuição dos dados. O algoritmo *SMOTE* gera um novo ponto de dado ao selecionar um ponto em uma linha que conecta uma amostra de classe minoritária escolhida aleatoriamente e um dos seus *k*-vizinhos mais próximos (ISLAM *et al*., 2022).

Em complemento, também foram utilizados nesta abordagem o algoritmo *Grid Search*, que segundo Velandia-Cardenas *et al*. (2021) é importante para estimar os melhores parâmetros do algoritmo *Random Forest* e o método de *k-fold Cross Validation* qu e é uma técnica de validação em que o conjunto de dados é dividido em *k* partes, usando *k–1* para treino e 1 para teste, repetindo o processo *k* vezes e calculando a média dos resultados, ao invés de apenas memorizar os exemplos que já conhece, como por exemplo no caso de separar as bases em conjuntos de dados de teste e treinamento fixos, além de retornar valores médios das métricas avaliadas e reduzir o *overfitting*. (RAGHUWANSHI; SHUKLA, 2019a), (FONTES *et al*., 2022). Um *script* em *Python* (Anexo B) foi desenvolvido e executado no ambiente remoto *Google Colab*[9] e os *steps* (passos) de todo este processo são detalhados a seguir e ilustrados na Figura 12.

Figura 12 - Fluxograma da abordagem de Classificação.



[9] https://colab.research.google.com/

1. **Importação e Análise de Dados**: As bibliotecas *pandas*, *numpy* e *seaborn* são importadas para permitir a leitura, manipulação e visualização dos dados, com a base de dados sendo carregada a partir de um arquivo *CSV* por meio do pandas, que a armazena em um *DataFrame*. Cada um dos 32 arquivos *CSV* que contém as bases sintéticas foi carregado e processado separadamente.

---

2. **Extração de Vetores:** No código, os vetores de entrada (X) e de saída (Y) são extraídos do *DataFrame*. O vetor Y é formado pela última coluna e o vetor X pelas demais. Esses vetores serão utilizados para análises subsequentes e construção de modelos.

---

3. **Análise de Separação entre Classes:** Neste *step*, o código aplica a técnica de Análise de Componentes Principais (PCA) para reduzir as dimensões do conjunto de dados a duas componentes principais utilizando a classe *PCA* da biblioteca *sklearn.decomposition* (apenas para as bases com 10 *Features*). Em seguida, é gerado um gráfico de dispersão (*scatter plot*), que visualiza a separação entre as classes, permitindo avaliar visualmente a distribuição das classes no espaço bidimensional. A Figura 13 exibe todos os *scatter plots* gerados identificando-os pela coluna **#** da Tabela 10.

---

4. **Aplicação do *SMOTE***: Neste *step* do código, a classe *SMOTE* da biblioteca *imblearn.over_sampling* é utilizada para aumentar a quantidade de amostras da classe minoritária no conjunto de dados, equilibrando assim a distribuição das classes. Após a aplicação do *SMOTE*, a distribuição das classes é verificada para confirmar o balanceamento dos dados.

---

5. ***Tuning* de Parâmetros com *Grid Search***: O código realiza um ajuste fino de hiperparâmetros (*criterion, max_depth, min_samples_leaf, min_samples_split*, e *n_estimators*) para um classificador *RF - Random Forest* (classe *RandomForestClassifier* da biblioteca *sklearn.ensemble*) utilizando o método de *Grid Search* (classe *GridSearchCV* da biblioteca *sklearn.model_selection)*. Especificamente, ele explora várias combinações de parâmetros, como o critério de divisão, o número de estimadores, a profundidade máxima das árvores, entre outros, para identificar a configuração que maximiza a métrica de *F1-Score* (funções *make_scorer* e *f1_score* da biblioteca *sklearn.metrics*, de acordo com

Velandia-Cardenas *et al*. (2021).

---

6. **Treinamento do Classificador *RF***: Aqui o classificador *RF* é treinado utilizando os melhores hiperparâmetros obtidos na etapa 5. O treinamento é realizado sobre os dados balanceados gerados pela etapa 4, permitindo que o modelo aprenda a distinguir entre as classes de forma mais eficaz, especialmente em cenários de desbalanceamento extremo.

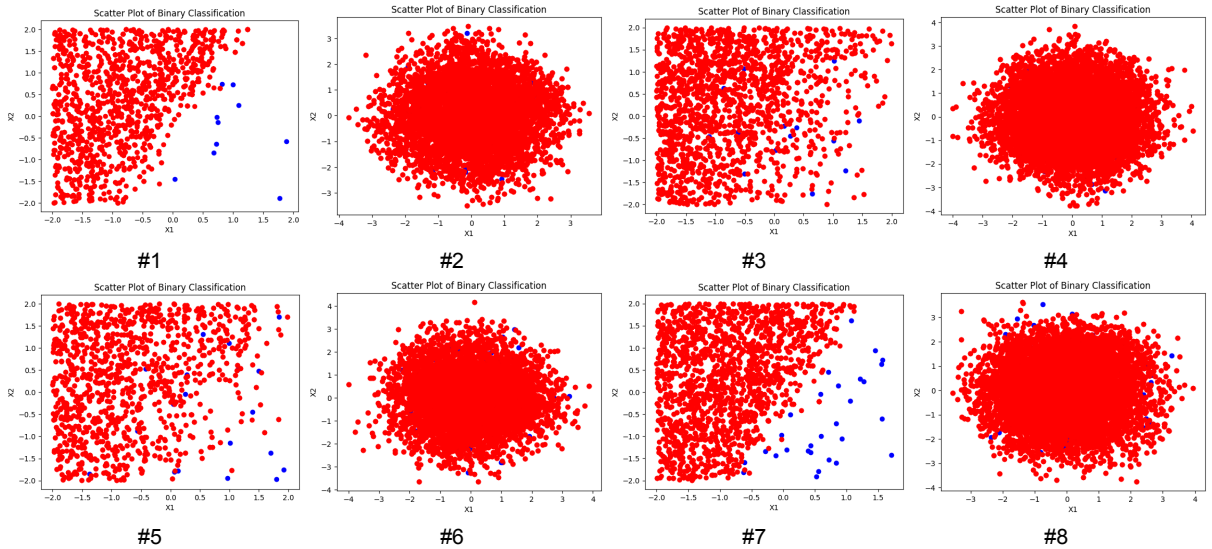
---

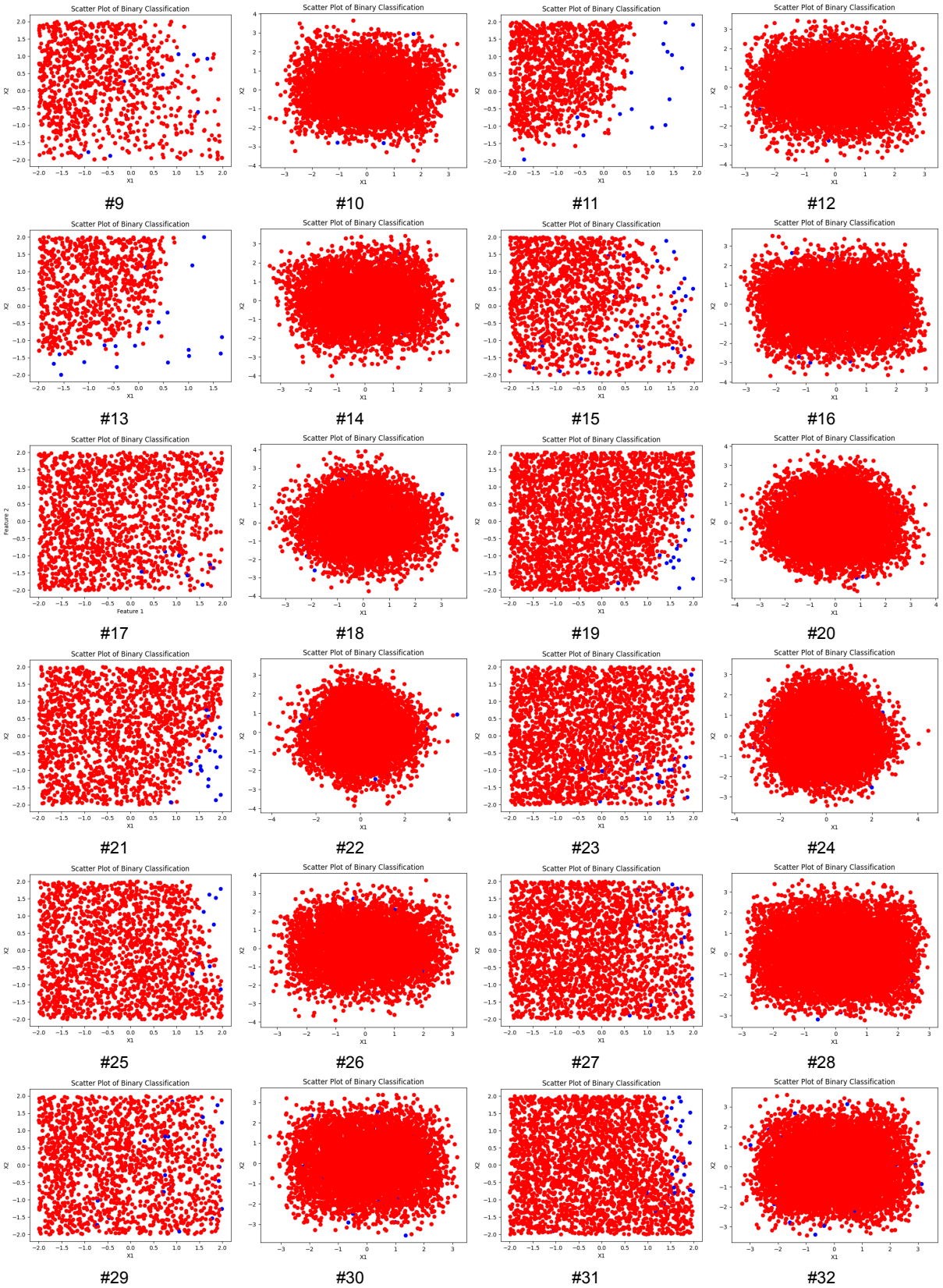
7. **Validação Cruzada com *k-fold***: Na etapa de validação cruzada com *K-Fold*, o desempenho do classificador RF é avaliado. O conjunto de dados é dividido em 10 partes (*folds*), e o modelo é treinado e testado 30 vezes, cada vez utilizando um fold diferente para teste e os outros para treino. Foi utilizada a classe *KFold* e a função *cross_val_score* da biblioteca *sklearn.model_selection*. De acordo com Zhang e Qin (2022), utilizar essa abordagem garante que o modelo seja avaliado de maneira robusta, reduzindo a variação nos resultados e proporcionando métricas de desempenho confiáveis, que neste experimento são: *ACC, PPV, TPR, F1-Score, AUC-ROC e G-Mean*.

---

8. **Análise do Resultados:** Neste *step* final, as métricas de desempenho obtidas durante a validação cruzada são compiladas em um *DataFrame* para análise estatística. São calculadas as médias, os desvios padrão e o coeficiente de variação das métricas, permitindo uma avaliação quantitativa da variabilidade e da confiabilidade do modelo.

Figura 13 - *Scatter plots* das bases sintéticas.





As Análises Estatísticas dos Dados e as Conclusões e Recomendações serão apresentadas e discutidas nas próximas seções.

4 RESULTADOS E DISCUSSÕES

4.1 Modelos de Classificação

Primeiramente serão apresentados resultados obtidos pelas métricas com a obtenção dos modelos de Classificação Binária. Com exceção da métrica Elapsed Time, que mediu o tempo gasto para a geração de cada base sintética, os valores das métricas **ACC, PPV, TPR, F1-Score, AUC-ROC e G-Mean** são exibidos contendo a média (*mean*), desvio padrão (*std*) e coeficiente de variação (*coef_var*) obtidos através da técnica de Validação Cruzada. A Tabela 11 apresenta os resultados.

O valor médio de cada métrica varia entre 0 e 1, ou seja, quanto mais próximo de 1 melhor é a capacidade preditiva do modelo. Contudo, em cenários de Desbalanceamento Extremo, analisar apenas uma métrica isoladamente pode levar a uma compreensão enganosa acerca do desempenho real. Considerar as métricas em conjunto, torna a análise dos modelos mais consistente e robusta.

No contexto de modelos de classificação, uma baseline ou análise ingênua refere-se a um modelo simples que serve como ponto de referência para avaliar a performance de modelos mais complexos. Em um cenário de dados Extremamente Desbalanceados, como o experimento apresentado neste artigo, uma análise ingênua poderia ser um modelo que sempre prevê a classe majoritária (neste caso, eventos falsos). Por exemplo, de acordo com os níveis definidos para o Fator C - Extreme Unbalance Percentage (onde apenas 1% ou 0,5% dos eventos são positivos), um modelo que sempre prevê "falso" teria uma precisão de 99% ou 99,5%. No entanto, esse modelo não seria útil para identificar os eventos positivos, que são a classe minoritária e de maior interesse. Portanto, a *baseline* apenas ajuda a entender o mínimo que um modelo deve alcançar para ser considerado significativo. A partir dessa *baseline*, as métricas são utilizadas para avaliar se modelos mais sofisticados estão realmente capturando a complexidade dos dados e melhorando a detecção das classes minoritárias. São apresentados em negrito na Tabela 11, todos os valores médios das métricas que ultrapassaram a baseline esperada para cada base de dados.

Do total de 32 bases sintéticas em apenas 6 (**18,75%**) a abordagem de Classificação aplicada não conseguiu obter ao menos um valor superior ao *baseline* em alguma das métricas. São bases que apresentam um padrão de Desbalanceamento Extremo desafiador e suas características (fatores e níveis) em comum serão discutidas e analisadas na seção subsequente.

Tabela 11 - Resultados da Classificação (métricas).

#	Elapsed Time	Recall (TPR)			AUC-ROC			F1-Score			Precision (PPV)			Accuracy (ACC)			G-Mean		
		mean	*std*	*coef_var*	*mean*	*std*	*coef_var*	*mean*	*std*	*coef_var*	*mean*	*std*	*coef_var*	*mean*	*std*	*coef_var*	*mean*	*std*	*coef_var*
17	0,069	0,9845	0,0015	0,1481	0,9842	0,0008	0,0812	0,9627	0,0014	0,1469	0,9421	0,0016	0,1656	0,9619	0,0013	0,1371	0,9616	0,0016	0,1612
15	0,146	0,9696	0,0022	0,2229	**0,9937**	0,0006	0,0578	0,9700	0,0016	0,1628	0,9708	0,0026	0,2672	0,9706	0,0018	0,1896	0,9702	0,0013	0,1319
4	0,599	**0,9999**	0,0001	0,0061	**1,0000**	0,0000	0,0005	**0,9999**	0,0001	0,0082	**0,9997**	0,0002	0,0177	**0,9999**	0,0001	0,0088	**0,9998**	0,0001	0,0085
7	0,158	**0,9946**	0,0007	0,0745	**0,9995**	0,0002	0,0228	**0,9951**	0,0006	0,0597	**0,9957**	0,0008	0,0800	**0,9951**	0,0006	0,0575	**0,9951**	0,0005	0,0543
10	0,407	**0,9999**	0,0001	0,0091	**1,0000**	0,0000	0,0001	**0,9997**	0,0002	0,0165	**0,9994**	0,0003	0,0335	**0,9997**	0,0002	0,0188	**0,9997**	0,0002	0,0181
8	0,557	**0,9993**	0,0002	0,0227	**1,0000**	0,0000	0,0015	**0,9988**	0,0003	0,0270	**0,9982**	0,0004	0,0448	**0,9988**	0,0002	0,0195	**0,9988**	0,0002	0,0223
32	0,589	**0,9992**	0,0002	0,0202	**1,0000**	0,0000	0,0010	**0,9968**	0,0003	0,0266	**0,9943**	0,0005	0,0550	**0,9968**	0,0002	0,0182	**0,9968**	0,0002	0,0245
6	0,445	**0,9999**	0,0001	0,0110	**1,0000**	0,0000	0,0005	**0,9995**	0,0002	0,0175	**0,9992**	0,0003	0,0325	**0,9995**	0,0001	0,0140	**0,9995**	0,0002	0,0181
27	0,126	0,9898	0,0011	0,1145	0,9853	0,0009	0,0867	0,9636	0,0011	0,1124	0,9396	0,0017	0,1841	0,9627	0,0013	0,1316	0,9624	0,0012	0,1216
14	0,407	**0,9990**	0,0002	0,0187	**1,0000**	0,0000	0,0011	**0,9983**	0,0004	0,0373	**0,9977**	0,0005	0,0539	**0,9983**	0,0003	0,0315	**0,9983**	0,0003	0,0307
12	0,529	**0,9998**	0,0001	0,0097	**1,0000**	0,0000	0,0004	**0,9993**	0,0002	0,0167	**0,9988**	0,0004	0,0357	**0,9993**	0,0002	0,0183	**0,9993**	0,0002	0,0190
30	0,406	**0,9996**	0,0002	0,0176	**1,0000**	0,0000	0,0002	**0,9985**	0,0002	0,0233	**0,9977**	0,0003	0,0340	**0,9985**	0,0002	0,0210	**0,9986**	0,0002	0,0180
3	0,124	0,9569	0,0012	0,1246	0,9926	0,0009	0,0890	0,9686	0,0020	0,2092	0,9805	0,0028	0,2814	0,9691	0,0018	0,1823	0,9688	0,0019	0,1969
25	0,087	**0,9972**	0,0010	0,1023	**0,9968**	0,0005	0,0532	0,9877	0,0006	0,0614	0,9784	0,0011	0,1125	0,9877	0,0006	0,0588	0,9876	0,0006	0,0622
24	0,548	**0,9993**	0,0001	0,0119	**1,0000**	0,0000	0,0003	**0,9987**	0,0001	0,0142	**0,9981**	0,0003	0,0259	**0,9987**	0,0002	0,0183	**0,9987**	0,0001	0,0118
2	0,496	**0,9998**	0,0001	0,0120	**1,0000**	0,0000	0,0008	**0,9994**	0,0002	0,0205	**0,9990**	0,0004	0,0446	**0,9994**	0,0002	0,0237	**0,9994**	0,0002	0,0218
19	0,187	0,9922	0,0012	0,1166	**0,9971**	0,0005	0,0470	0,9891	0,0007	0,0725	0,9864	0,0009	0,0908	0,9888	0,0007	0,0683	0,9891	0,0008	0,0785
16	0,612	**0,9998**	0,0001	0,0133	**1,0000**	0,0000	0,0003	**0,9994**	0,0002	0,0182	**0,9990**	0,0004	0,0383	**0,9994**	0,0002	0,0196	**0,9995**	0,0002	0,0168
31	0,210	0,9789	0,0010	0,1056	**0,9935**	0,0004	0,0359	0,9736	0,0009	0,0893	0,9691	0,0012	0,1282	0,9735	0,0008	0,0815	0,9736	0,0009	0,0928
13	0,087	0,9692	0,0018	0,1853	**0,9982**	0,0004	0,0397	0,9786	0,0015	0,1501	0,9884	0,0019	0,1946	0,9787	0,0015	0,1522	0,9788	0,0013	0,1357
22	0,427	**0,9986**	0,0003	0,0286	**0,9998**	0,0000	0,0048	**0,9957**	0,0003	0,0344	**0,9928**	0,0006	0,0558	**0,9957**	0,0004	0,0359	**0,9957**	0,0003	0,0272
29	0,115	0,9416	0,0027	0,2858	0,9679	0,0009	0,0947	0,9227	0,0016	0,1752	0,9039	0,0024	0,2692	0,9210	0,0020	0,2183	0,9208	0,0018	0,2000
21	0,138	**0,9978**	0,0006	0,0581	**0,9968**	0,0004	0,0419	**0,9927**	0,0006	0,0567	0,9877	0,0008	0,0833	**0,9927**	0,0005	0,0524	**0,9927**	0,0006	0,0570
23	0,128	0,9589	0,0019	0,1996	0,9792	0,0007	0,0700	0,9414	0,0015	0,1611	0,9248	0,0021	0,2302	0,9405	0,0013	0,1360	0,9402	0,0015	0,1626
9	0,086	0,9689	0,0019	0,1953	0,9926	0,0007	0,0715	0,9707	0,0017	0,1777	0,9725	0,0025	0,2604	0,9709	0,0018	0,1880	0,9708	0,0016	0,1675
1	0,107	**0,9999**	0,0002	0,0205	**1,0000**	0,0000	0,0001	**0,9999**	0,0002	0,0213	**0,9998**	0,0003	0,0282	**0,9999**	0,0002	0,0158	**0,9998**	0,0002	0,0232
11	0,151	0,9949	0,0009	0,0869	**0,9998**	0,0002	0,0174	**0,9966**	0,0005	0,0517	**0,9981**	0,0006	0,0567	**0,9965**	0,0005	0,0528	**0,9966**	0,0005	0,0505
18	0,459	**0,9999**	0,0001	0,0113	**1,0000**	0,0000	0,0001	**0,9995**	0,0001	0,0111	**0,9991**	0,0002	0,0155	**0,9995**	0,0001	0,0144	**0,9995**	0,0001	0,0133
26	0,382	**1,0000**	0,0001	0,0082	**1,0000**	0,0000	0,0000	**0,9986**	0,0002	0,0236	**0,9972**	0,0005	0,0457	**0,9986**	0,0002	0,0214	**0,9986**	0,0002	0,0198
28	0,560	**0,9999**	0,0001	0,0085	**1,0000**	0,0000	0,0000	**0,9996**	0,0002	0,0151	**0,9993**	0,0002	0,0226	**0,9996**	0,0001	0,0115	**0,9996**	0,0001	0,0107
20	0,511	**0,9997**	0,0001	0,0125	**1,0000**	0,0000	0,0004	**0,9984**	0,0002	0,0170	**0,9972**	0,0003	0,0308	**0,9984**	0,0002	0,0166	**0,9984**	0,0002	0,0161
5	0,120	0,9582	0,0023	0,2434	**0,9921**	0,0009	0,0932	0,9683	0,0022	0,2284	0,9803	0,0035	0,3584	0,9690	0,0022	0,2220	0,9690	0,0018	0,1894

4.2 Análises Estatísticas dos Dados

De acordo com Montgomery (2012) os métodos estatísticos adicionam objetividade ao processo de tomada de decisão. Técnicas estatísticas, aliadas a um bom conhecimento do processo aliado ao bom senso, via de regra levam a conclusões sólidas acerca do objeto de estudo.

Levando em conta que o resultado de cada um das seis métricas de avaliação dos modelos de Classificação Binária é composto por média, desvio padrão e coeficiente de variação, totalizando 18 respostas. Todavia, as análises estatísticas do experimento relativas à variável de resposta que mediu o tempo de execução computacional para a geração de cada uma das bases sintéticas (*Elapsed Time*), serão apresentadas separadamente na Tabela 12, devido a característica desta resposta.

Tabela 12 - Análises da Resposta *Elapsed Time*

Elapsed Time
1. Modelo Geral: O Modelo tem um valor de F de 104,57 e um valor de P de 0,000, o que indica que o modelo geral é significativo e que há uma variação explicada pelos fatores. **2. Efeitos Lineares Relevantes:** *Features*: *P-value* de 0,000 (F = 2000,47), indicando uma influência extremamente significativa. *Sampling*: *P-value* de 0,000 (F = 129,01), também com impacto significativo. *Error*: *P-value* de 0,024 (F = 7,01) o que indica um impacto significativo, embora menor. **3. Interações Significativas de 2 fatores:** *Features*Sampling:* P-valor de 0,001 (F = 24,70), uma interação altamente significativa. *Sampling*Y function*: P-valor de 0,013 (F = 8,98), também significativo. **4. Ajuste:** S = 0,0232993; R-sq = 99,55%; R-sq(adj) = 98,59%; R-sq(pred) = 95,36%. O modelo tem um ótimo ajuste, explicando quase toda a variabilidade observada nos dados, com boa capacidade preditiva em relação ao tempo de execução e geração das bases.

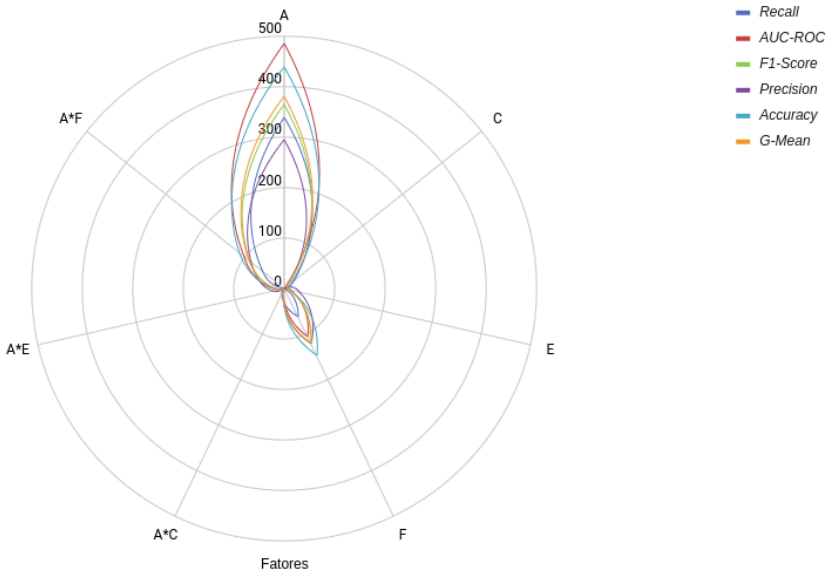
Os principais insights sobre os efeitos de diferentes fatores e interações no experimento serão discutidos com base na Análise de Variância (ANOVA), com foco na interpretação dos fatores mais relevantes e interações estatisticamente significativas. Para facilitar a interpretação, apenas os fatores e interações de segunda ordem com *F-Value* e *P-Value* significativos para cada resposta são apresentados. Esses fatores, representados na forma de um mapa de calor, são exibidos na Tabela 13, destacando aqueles com os maiores valores para cada resposta pela soma dos resultados *F-Values* para média, desvio padrão (*std*) e coeficiente de variação (*coef_var*), indicando assim os fatores de alta significância estatística.

Tabela 14 - Fatores significativos (mapa de calor).

Fatores	Recall	AUC-ROC	F1-Score	Precision	Accuracy	G-Mean
[A]	339,59	486,68	363,93	295,79	439,53	380,58
[F]	61,96	105,91	119,76	121,66	147,28	120,49
[A]*[F]	52,01	102,4	87,20	82,32	113,01	90,03
[E]	0	12,11	9,21	30,23	9,54	9,61
[C]	24,02	0	6,76	6,66	19,23	6,28
[A]*[E]	0	11,97	6,84	23,75	7,20	7,28
[A]*[C]	10,52	0	0	0	0	0

Em complemento, a Figura 14, compara visualmente o impacto dos fatores nas métricas (ACC, PPV, TPR, F1-Score, AUC-ROC e G-Mean), permitindo a análise de desempenho de cada fator em múltiplas métricas simultaneamente através dos dados apresentados no mapa de calor.

Figura 14 - Fatores Significativos.



4.3 Conclusões e Recomendações

O Planejamento de Experimentos (*DoE*) em caráter exploratório é uma abordagem sistemática utilizada para investigar os efeitos de múltiplos fatores sobre uma ou mais respostas de interesse, em cenários onde o conhecimento prévio sobre o sistema ou processo é limitado. Nesse contexto, nosso trabalho teve como objetivo identificar variáveis relevantes e potenciais interações, bem como fornecer uma base inicial de informações para experimentos subsequentes mais detalhados. Ao explorar um conjunto de fatores simultaneamente, é possível obter insights sobre a estrutura do problema relativo ao Desbalanceamento Extremo, auxiliando na delimitação de hipóteses e na priorização de variáveis que terão maior impacto sobre os resultados. Esse tipo de planejamento é particularmente útil em fases preliminares de pesquisa, onde a exploração de combinações de fatores pode revelar padrões e relações ainda não conhecidas.

A fim de aprofundarmos mais nas análises dos Resultados da Classificação exibidos na Tabela 11, separamos as bases de dados sintéticas e suas características com resultados inferiores ao *baseline* e exibimos na Tabela 15.

Tabela 15 - Bases com performance inferior ao *baseline.*

Base	**Fatores e níveis**					
#	**A**	**B**	**C**	**D**	**E**	**F**
17	**2**	1000	0,5%	Linear Simples	0,8	**0.8**
27	**2**	1500	0,5%	Com interações	0,8	**0.8**
3	**2**	1500	0,5%	Linear Simples	0,5	**0.8**
29	**2**	1000	1%	Com interações	0,8	**0.8**
23	**2**	1500	1%	Linear Simples	0,8	**0.8**
9	**2**	1000	0,5%	Com interações	0,5	**0.8**

Pode-se observar que os fatores **A** e **F** apresentam um único nível em todas as bases sintéticas que obtiveram os piores resultados: **A (nível -)[=2]** e **F (nível +)[=0,8]**. Este comportamento é coerente com as análise estatística dos dados por meio do *DoE*, pois de acordo com a Tabela 14 (mapa de calor) e o gráfico da radar da Figura 14, os fatores **A** e **F** e interação **A*F** apresentam a maior significância em relação aos demais, tendo o fator **A** isoladamente o maior valor de significância estatística entre todos.

De tal modo, em caráter secundário, temos o fatores **C** (maior incidência do **nível -**) e **E** (maior incidência do **nível +**) e a interação de ambos com A (**A*C** e **A*E**) apresentando um impacto com menos significância, mas que também precisam ser analisados para um entendimento mais completo no fenômeno.

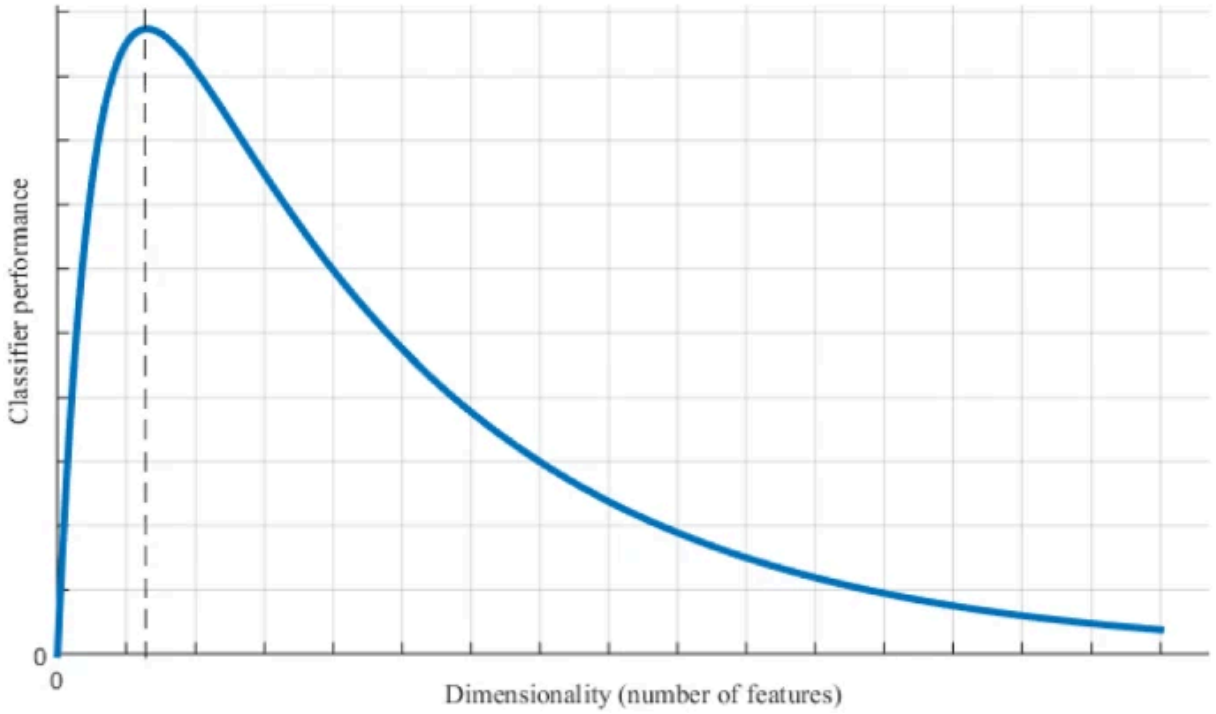
4.3.1 Análises sobre Dimensionalidade - Fator A

A redução da dimensionalidade é uma abordagem muito utilizada em machine learning para simplificar modelos, melhorar a eficiência computacional e reduzir o risco de *overfitting*. Segundo Abdulrauf Sharifai e Zainol (2020), mais especificamente, a classificação de conjuntos de dados desequilibrados, torna-se ainda mais complexa conforme o número de features aumenta, contexto este que pode prejudicar o desempenho dos algoritmos. Entretanto, como visto nos resultados deste trabalho, um baixo número de features também pode ser tão prejudicial ao modelo quanto o alto número dimensional.

Para que a redução tenha resultados positivos é importante que o *trade-off* entre a diminuição da capacidade descritiva de um dado problema em relação ao déficit de performance métrico, uma vez que em cenários de desbalanceamento extremo, cada 0,01 a menos nas principais métricas de avaliação acarreta significativamente a capacidade preditiva dos modelos.

Na Figura 15, o desempenho de um dado classificador demonstra uma melhora inicial com o aumento da dimensionalidade. No entanto, à medida que a dimensionalidade continua a aumentar além de um ponto ótimo, o desempenho do classificador começa a se degradar (JIA *et a*l., 2022). O baixo número de *features* (natural ou processado) tem efeito negativo similar a alta dimensionalidade.

Figura 15 - Tendência do desempenho do classificador (JIA *et a*l., 2022).



Todavia existem problemas que possuem por natureza uma quantidade muito reduzida de features, como por exemplo os casos abordados nos trabalhos de (RAGHUWANSHI; SHUKLA, 2019a) e (HU *et al.,* 2022). Neste contexto, as abordagens mais adequadas tendem a focar em aumentar a complexidade do espaço de características ou extrair mais informações das poucas features disponíveis. Neste cenário, algumas estratégias e técnicas incluem:

a. **Aumentar a Complexidade das *Features* via *Feature Engineering***: De acordo com Verdonck et al. (2021), *Feature Engineering* pode ser definida como a manipulação inteligente dos dados, aproveitando o viés intrínseco das técnicas de *machine learning* em nosso benefício, idealmente melhorando tanto a precisão quanto a interpretabilidade simultaneamente. Ela tem como objetivo projetar atributos significativos por meio de duas abordagens principais: **i**. ajustando características existentes através de diversas transformações (como logarítmicas e exponenciais por exemplo); **ii**. extraindo e criando novos atributos relevantes a partir de diferentes fontes (tais como combinações entre *features*: multiplicações, somas, diferenças ou termos polinomiais: quadrados, cubos, etc.) — um processo frequentemente chamado de "*featurização*".
b. **Modelos de Kernel: i**. SVM (*Support Vector Machines*) com *kernels* não lineares (ex.: *kernel* radial, polinomial), permite projetar os dados em um espaço de dimensão maior sem necessidade explícita de gerar novas *features*. Conforme Ghosh *et al*. (2019), este algoritmo pode ser usado para gerar um limite de decisão não linear usando a função kernel; **ii**. *Gaussian Processes* utilizam *kernels* para mapear os dados para um espaço de dimensão maior, possibilitando a modelagem de relações complexas entre as *features*, abordagem que pode ser vista no trabalho de (RAGHUWANSHI; SHUKLA, 2019b).
c. **Outras abordagens**: **i**. *Data Augmentation* pode ser usado em casos com poucas *features* e dados limitados. Isso é comum em dados de imagem, mas também pode ser adaptado para outros tipos de dados. De acordo com Maharana *et al*. (2022) as técnicas de *Data Augmentation* geram dados para modelos de aprendizado de máquina, diminuindo a dependência dos dados de treinamento e melhorando o desempenho do modelo; **ii**. Técnicas de *Embeddings* para transformar dados, especialmente categóricos ou textuais, em vetores de números que capturam relações ou significados importantes entre esses dados, como exemplo: *Word Embeddings* (SELVA BIRUNDA; KANNIGA DEVI, 2021) e *Entity Embeddings* (PIAO, 2021); **iii**. Técnicas Bayesianas, conforme Lantz (2019), em casos de poucas *features*, modelos que adotam uma abordagem probabilística, como redes

Bayesianas ou regressão Bayesiana, podem ser vantajosos em cenários de alta variabilidade ou dados limitados, pois conseguem incorporar incertezas nas estimativas.

4.3.2 Análises sobre o Termo de Erro da Função - Fator F

O termo de erro representa a variabilidade nos dados que o modelo não consegue explicar, ou seja, a porção da variação não atribuível às variáveis independentes. Este termo, conforme Montgomery (2012), é denominado erro aleatório. Neste experimento, o erro foi manipulado artificialmente para a geração das bases sintéticas, com o objetivo de tentar simular o comportamento de fenômenos e comportamentos da natureza.

Assume-se que o erro possui média igual a zero $(E(\epsilon) = 0)$, com variações percentuais (controladas pelos níveis) sobre o desvio padrão (σ) dos valores de y. Para o nível inferior (-), aplica-se 20% do valor do desvio padrão de y ao termo de erro, enquanto para o nível superior (+) aplica-se 80%. Dessa forma, o termo de erro aleatório (ϵ) segue uma distribuição normal com média zero e desvio padrão escalado de acordo com esses percentuais, em relação ao desvio padrão de y. De acordo com a Tabela 7, todas as bases que não alcançaram o *baseline* foram geradas com o **nível +** de F (0,8), ou seja, quanto mais alta a variação do termo, menor fica a precisão dos modelos de classificação.

Este fator permite que o modelo incorpore a variabilidade aleatória, refletindo condições mais realistas e promovendo uma análise robusta da resposta frente aos níveis especificados do termo de erro, o que torna as bases sintéticas geradas mais desafiadoras e próximas do mundo real. As Figuras 16 e 17 demonstram como os efeitos do fator F impactam significativamente o comportamento da variável dependente (y).

Figura 16 - Exemplo de resposta com termo de erro com **nível -** .

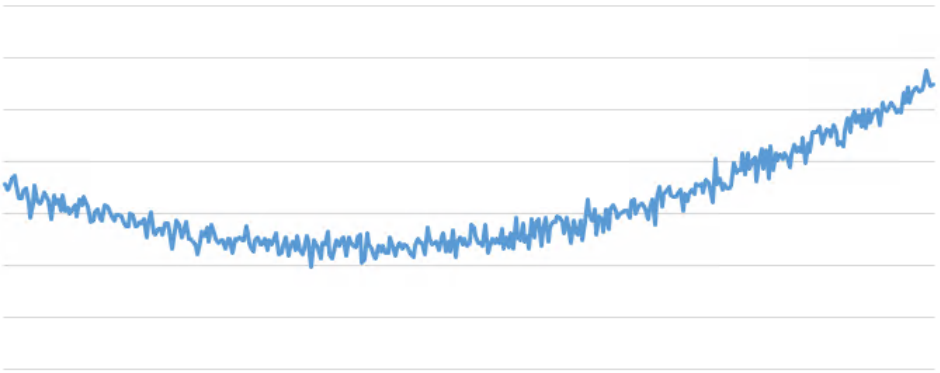
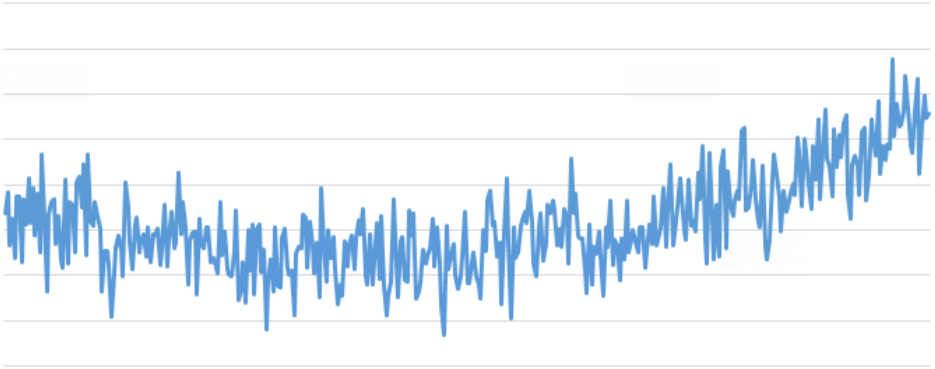


Figura 17 - Exemplo de resposta com termo de erro com **nível +** .



Para lidar com cenários similares a estes, algumas estratégias e técnicas discutidas na análise anterior também podem ser aplicadas, tais como: *Feature Engineering* e *Transfer Learning* e outras abordagens discutidas na literatura já foram aplicadas para este experimento: Métodos de *Ensemble* e *Bootstrap Aggregating - Bagging* (*Random Forest*), Técnicas de balanceamento (*SMOTE*), *Cross Validation* (*k-fold)* e *Hyperparameter Tuning* (*grid search).* Entretanto, outras abordagens complementares podem ser avaliadas, dentre as quais destacam-se:

a. **Regularização**: abordagens que adicionam penalidades que limitam a magnitude dos coeficientes, reduzindo a complexidade do modelo, tornando-o menos suscetível a oscilações aleatórias nos dados e menos propenso a overfitting, uma vez que minimiza o impacto do ruído na generalização do modelo (TIAN; ZHANG, 2022).
b. **Adoção de Modelos de *Deep Learning***: adoção de abordagens de redes neurais profundas, como por exemplo as baseadas na arquitetura de Redes Neurais Convolucionais (*CNNs*) para imagens, como visto do trabalho de Xia *et al*. (2021), podem capturar padrões complexos que métodos tradicionais não conseguem.
c. **Métodos de Detecção e Tratamento de *Outliers* e Anomalias:** abordagens que permitem identificar e tratar observações que fogem da distribuição esperada, seja através de remoção ou tratamento apropriado. Como por exemplo: *Isolation Forest,* utilizado no trabalho de Fontes *et al*. (2022) ou *Local Outlier Factor (LOF),* uma das técnicas mais conhecidas para detecção de *outliers* locais baseada em densidade (ALGHUSHAIRY *et al*., 2020)..
d. **Outras abordagens: i**. *Noise Robust Models*; **ii.** Ajustes no Pipeline de Validação (*Holdout* Dinâmico e Validação Estratificada) e Testes com Conjuntos Independentes;

Cada uma dessas alternativas fornece ao modelo um potencial adicional de adaptação aos dados, minimizando a influência do ruído e da variabilidade aleatória. A combinação de técnicas avançadas, associada a uma análise minuciosa dos dados, pode

contribuir para melhorar os resultados e aumentar a precisão e a robustez dos modelos de classificação em contextos desafiadores, como as bases sintéticas apresentadas em nosso trabalho.

4.3 Considerações Finais

Esta pesquisa ofereceu uma abordagem sistemática e controlada para a geração de dados sintéticos extremamente desbalanceados, com foco em cenários de classificação binária, cenário este que é uma condição recorrente em aplicações críticas de *Machine Learning*.

Baseado no *rodamap* de *DoE*, um processo robusto para a criação de bases de dados sintéticas foi desenvolvido com vistas a atender aos requisitos de representatividade e variabilidade para avaliação de modelos de classificação.

Os resultados obtidos demonstram que a combinação do *ensemble Random Forest* e a técnica de pré-processamento *SMOTE* proporcionam uma capacidade preditiva aprimorada em cenários de desbalanceamento extremo, conforme o levantamento teórico apresentado pela RSL no Capítulo 2. Os resultados das métricas avaliadas: *TPR, PPV, AUC-ROC, Accuracy, F1-Score e G-MEAN*, fornecem uma base confiável para validar a aplicação experimental. Estes resultados, particularmente no que tange à superação da *baseline* em 81,25% dos casos analisados, destacam a relevância da abordagem proposta em relação às bases geradas.

A análise estatística detalhada, amparada pelo uso de técnicas como a ANOVA, permitiu uma interpretação aprofundada dos fatores e interações mais significativos, destacando o impacto dos fatores **A (*features*)** e **F (erro)** no desempenho dos modelos de classificação. Esses *insights* contribuem para um entendimento mais claro das características intrínsecas de conjuntos de dados desbalanceados, assim como para a identificação de estratégias que possam mitigar os desafios associados a essas características.

Em síntese, este trabalho não apenas valida a viabilidade da proposta de abordagem para geração de dados sintéticos em contextos de desbalanceamento extremo, mas também oferece uma estrutura metodológica replicável que pode ser adotada e aprimorada em estudos futuros.

4.4 Limitações Experimentais do Trabalho

O delineamento experimental deste estudo baseia-se na geração de bases sintéticas por meio de DoE, com variação controlada de fatores como grau de desbalanceamento, separabilidade das classes e tamanho da amostra. O objetivo central é avaliar a influência desses fatores sobre métricas de desempenho de um pipeline de classificação, por meio de Análise de Variância (ANOVA). As métricas analisadas representam respostas quantitativas ao experimento.

Para mensurar essas respostas, empregou-se um pipeline de classificação que utiliza o algoritmo SMOTE antes da divisão treino–teste, prática que, do ponto de vista de machine learning, é reconhecida como suscetível a vazamento de dados. Essa escolha foi deliberada, pois o propósito deste trabalho não é estimar o desempenho real do classificador em produção, pois a RSL já ratifica o potencial de performance das abordagens de classificação adotadas. Contudo, investigar de forma controlado como os fatores do DoE impactam as métricas de avaliação dentro de um mesmo contexto de modelagem.

O uso do SMOTE global tende a inflar os valores absolutos das métricas. No entanto, como o procedimento é idêntico em todos os cenários experimentais, o viés introduzido atua de maneira sistemática. Assim, as comparações entre níveis e fatores — essência da ANOVA — permanecem válidas, permitindo identificar diferenças relativas e efeitos principais com rigor estatístico. Em outras palavras, a ANOVA avalia contrastes entre tratamentos e não depende de valores absolutos livres de viés, desde que o processo de avaliação seja consistente entre os ensaios (MONTGOMERY, 2019).

É importante, todavia, explicitar a limitação: os resultados da ANOVA devem ser interpretados dentro do contexto específico do pipeline adotado, não como estimativas de desempenho generalizáveis a ambientes de produção. A presente análise, portanto, quantifica a sensibilidade das métricas ao delineamento dos fatores do DoE em um cenário controlado, evidenciando que a validade estatística da comparação entre fatores não é comprometida pelo viés uniforme do SMOTE.

4.5 Trabalhos Futuros

Este trabalho teve seu escopo focado em problemas de Classificação Binária em bases de dados Extremamente Desbalanceadas com objetivo de encontrar as abordagens mais eficientes através de estudos primários e por meio de DoE, propor uma abordagem

inovadora para geração de bases sintéticas controladas. Porém ainda há neste contexto próximo outras oportunidades para geração de conhecimento. Podemos destacar:

i. problemas relacionados à classificação multiclasses, ou seja, onde temos mais do que dois rótulos de saída. Embora o Desbalanceamento Extremo também seja um problema nestes casos, existem especificidades e outros desafios quanto à escolha das abordagens mais eficientes e ajustadas para lidar com estes problemas;

ii. outros fatores que também influenciam na qualidade dos modelos: o número de características (*features*) e amostras utilizadas. A seleção de características é crucial para o desenvolvimento de modelos de Aprendizagem de Máquina. Escolher um subconjunto relevante de características ajuda a reduzir a complexidade do modelo, evitando o *overfitting* e garantindo a capacidade de generalização. Além disso, a quantidade e qualidade das amostras de treinamento também impactam significativamente a capacidade do modelo de aprender e generalizar padrões. Portanto, pode ser acrescentado ao problema de Desbalanceamento Extremo, estudos que analisem a performance das abordagens em cenários com número elevado de características (*features*) e com número mais reduzido de amostras;

iii. o desenvolvimento de uma interface gráfica amigável (user-friendly) para o banco de dados que reúne os achados da Revisão Sistemática da Literatura. Tal ferramenta permitiria acesso mais simplificado e intuitivo às informações catalogadas, favorecendo consultas e análises por pesquisadores e profissionais interessados em identificar as abordagens de melhor performance em diferentes cenários de desbalanceamento extremo;

iv. a ampliação das análises realizadas com base no Design de Experimentos (DoE), incorporando o uso de métodos multivariados. Essa abordagem poderia oferecer uma compreensão mais aprofundada dos fatores e de suas interações, permitindo avaliar de forma mais abrangente os conjuntos de resultados obtidos e suas implicações práticas.

v. adoção de *pipelines* alternativas de classificação binária para outras análises e *benchmarks.*

REFERÊNCIAS

ABDULRAUF SHARIFAI, G.; ZAINOL, Z. Feature Selection for High-Dimensional and Imbalanced Biomedical Data Based on Robust Correlation Based Redundancy and Binary Grasshopper Optimization Algorithm. **Genes**, Basel, v. 11, n. 7, p. e717, jul. 2020. DOI: 10.3390/genes11070717.

ALGHUSHAIRY, O. *et al*. A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams. **Big Data and Cognitive Computing**, Basel, v. 5, n. 1, p. e1, 2020. DOI: 10.3390/bdcc5010001.

ALSOBHI, H. A. *et al*. Blockchain-based micro-credentialing system in higher education institutions: Systematic literature review. **Knowledge-Based Systems**, v. 265, p. 110238, abr. 2023.

BISTROŃ, M.; PIOTROWSKI, Z. Comparison of Machine Learning Algorithms Used for Skin Cancer Diagnosis. **Applied Sciences**, Basel, v. 12, n. 19, p. e9960, 2022. DOI: 10.3390/app12199960.

BOKONDA, P. L.; OUAZZANI-TOUHAMI, K.; SOUISSI, N. Predictive analysis using machine learning: Review of trends and methods. *In*: INTERNATIONAL SYMPOSIUM ON ADVANCED ELECTRICAL AND COMMUNICATION TECHNOLOGIES (ISAECT), 2020, [S. l.]. **Anais** [...]. [S. l.]: IEEE, 2020. p. 1-6. DOI: 10.1109/isaect50560.2020.9523703.

CANBEK, G. *et al*. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. *In*: INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND ENGINEERING (UBMK), 2017, [S. l.]. **Anais** [...]. [S. l.]: IEEE, 2017. p. 811-816. DOI: 10.1109/ubmk.2017.8093539.

CHEN, Y.; SHAYILAN, A. Dictionary learning for multivariate geochemical anomaly detection for mineral exploration targeting. **Journal of Geochemical Exploration**, v. 235, p. 106958, abr. 2022.

CLASSIFICAÇÃO: curva ROC e AUC. [S. l.]: Google for Developers, 2022. Disponível em: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=pt-br. Acesso em: 9 fev. 2024.

DANGUT, M. D. *et al*. Application of deep reinforcement learning for extremely rare failure prediction in aircraft maintenance. **Mechanical Systems and Signal Processing**, v. 171, p.

108873, maio 2022.

DATAR, R.; GARG, H. **Hands-On Exploratory Data Analysis with R**: Become an expert in exploratory data analysis using R packages. Birmingham: Packt Publishing, 2019.

DEAN, A.; VOSS, D.; DRAGULJIĆ, D. **Design and Analysis of Experiments**. 2. ed. Cham: Springer, 2017.

FAN, C. *et al*. Quantitative assessments on advanced data synthesis strategies for enhancing imbalanced AHU fault diagnosis performance. **Energy and Buildings**, v. 252, p. 111423, dez. 2021.

FONTES, G. *et al*. An Empirical Study on Anomaly Detection Algorithms for Extremely Imbalanced Datasets. *In*: MAGLOGIANNIS, I.; MACINTYRE, J.; ILIADIS, L. (ed.). **Artificial Intelligence Applications and Innovations**. Cham: Springer, 2022. p. 85-95. (IFIP Advances in Information and Communication Technology, v. 649). DOI: 10.1007/978-3-031-08337-2_7.

GHOSH, S.; DASGUPTA, A.; SWETAPADMA, A. A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification. *In*: INTERNATIONAL CONFERENCE ON INTELLIGENT SUSTAINABLE SYSTEMS (ICISS), 2019, Palladam. **Anais** [...]. New York: IEEE, 2019. p. 24-28. DOI: 10.1109/ISS1.2019.8908018.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge: MIT Press, 2016.

HAO, S. *et al*. **Synthetic Data in AI**: Challenges, Applications, and Ethical Implications. [S. l.]: arXiv, 2024. Preprint. DOI: 10.48550/arXiv.2401.01629.

HU, J. *et al*. BBW: a batch balance wrapper for training deep neural networks on extremely imbalanced datasets with few minority samples. **Applied Intelligence**, v. 52, n. 6, p. 6723–6738, 2021.

IMBALANCED Data. [S. l.]: Google for Developers, 2025. Disponível em: https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data. Acesso em: 9 fev. 2025.

ISLAM, A. *et al*. KNNOR: An oversampling technique for imbalanced datasets. **Applied Soft Computing**, Amsterdam, v. 115, p. e108288, 2022. DOI: 10.1016/j.asoc.2021.108288.

JIA, W. *et al*. Feature dimensionality reduction: a review. **Complex & Intelligent Systems**,

Berlin, v. 8, n. 3, p. 2663-2693, 2022. DOI: 10.1007/s40747-021-00637-x.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, Washington, DC, v. 349, n. 6245, p. 255-260, jul. 2015. DOI: 10.1126/science.aaa8415.

KAUR, P.; GOSAIN, A. FF-SMOTE: A Metaheuristic Approach to Combat Class Imbalance in Binary Classification. **Applied Artificial Intelligence**, v. 33, n. 5, p. 420–439, 2019.

KAUR, P.; GOSAIN, A. Robust hybrid data-level sampling approach to handle imbalanced data during classification. **Soft Computing**, v. 24, n. 20, p. 15715–15732, 2020.

KITCHENHAM, B. *et al*. Systematic literature reviews in software engineering – A systematic literature review. **Information and Software Technology**, v. 51, n. 1, p. 7–15, jan. 2009.

LANTZ, B. **Machine Learning with R**: Expert techniques for predictive modeling. 3. ed. Birmingham: Packt Publishing, 2019.

LAQUEUR, H. S. *et al*. Machine Learning Analysis of Handgun Transactions to Predict Firearm Suicide Risk. **JAMA Network Open**, v. 5, n. 7, p. e2221041, 2022.

LI, Y. *et al*. Imbalanced text sentiment classification using universal and domain-specific knowledge. **Knowledge-Based Systems**, v. 160, p. 1–15, nov. 2018.

LU, C. *et al*. An improved weighted extreme learning machine for imbalanced data classification. **Memetic Computing**, v. 11, n. 1, p. 27–34, 2019.

LU, S. *et al*. Dynamic Weighted Cross Entropy for Semantic Segmentation with Extremely Imbalanced Data. *In*: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND ADVANCED MANUFACTURING (AIAM), 2019, [S. l.]. **Anais** [...]. [S. l.]: IEEE, 2019. p. 1-4. DOI: 10.1109/aiam48774.2019.00053.

MAHARANA, K.; MONDAL, S.; NEMADE, B. A review: Data pre-processing and data augmentation techniques. **Global Transitions Proceedings**, Amsterdam, v. 3, p. 91-99, 2022. DOI: 10.1016/j.gltp.2022.04.020.

MATIAS-PEREIRA, J. **Manual de metodologia da pesquisa científica**. [S. l.]: Grupo GEN, 2016.

MIAO, J.; ZHU, W. Precision–recall curve (PRC) classification trees. **Evolutionary Intelligence**, v. 15, n. 3, p. 1545–1569, 2021.

MINITAB. **Análise de DOE**. State College, [20--?]. Disponível em: https://support.minitab.com/pt-br/engage/help-and-how-to/tools/forms/form-tools/statistical-analysis/doe-analysis/. Acesso em: 30 jul. 2024.

MINITAB. **Experimentos fatoriais e fatoriais fracionados**. State College, [20--?]. Disponível em: https://support.minitab.com/pt-br/minitab/help-and-how-to/statistical-modeling/doe/supporting-topics/factorial-and-screening-designs/factorial-and-fractional-factorial-designs/. Acesso em: 30 jul. 2024.

MONTGOMERY, D. C. **Design and analysis of experiments**. 8. ed. Hoboken: John Wiley & Sons, 2012.

MONTGOMERY, D. C. **Design and analysis of experiments**. 10. ed. Hoboken: Wiley, 2019.

MURPHY, K. P. **Machine Learning**: A Probabilistic Perspective. Cambridge: MIT Press, 2012.

MURTAZA, H. *et al*. Synthetic data generation: State of the art in health care domain. **Computer Science Review**, v. 48, p. 100546, maio 2023. DOI: 10.1016/j.cosrev.2023.100546.

NOVIYANTO, A.; ABDULLA, W. H. Honey botanical origin classification using hyperspectral imaging and machine learning. **Journal of Food Engineering**, v. 265, p. 109684, jan. 2020.

PEREIRA, P. J. *et al*. A Comparison of Machine Learning Methods for Extremely Unbalanced Industrial Quality Data. *In*: MARREIROS, G. *et al*. (ed.). **Progress in Artificial Intelligence**. Cham: Springer, 2021. p. 561-572. (Lecture Notes in Computer Science, v. 12981).

PIAO, G. Scholarly Text Classification with Sentence BERT and Entity Embeddings. *In*: STOREY, V. C. *et al*. (ed.). **Conceptual Modeling**. Cham: Springer, 2021. p. 79-87. (Lecture Notes in Computer Science, v. 12584). DOI: 10.1007/978-3-030-65847-2_8.

RAGHUWANSHI, B. S.; SHUKLA, S. Classifying imbalanced data using ensemble of reduced kernelized weighted extreme learning machine. **International Journal of Machine Learning and Cybernetics**, v. 10, n. 11, p. 3071–3097, 2019a.

RAGHUWANSHI, B. S.; SHUKLA, S. Classifying imbalanced data using BalanceCascade-based kernelized extreme learning machine. **Pattern Analysis and Applications**, v. 23, n. 3, p. 1157–1182, 2019b.

REN, J. *et al*. Equalization ensemble for large scale highly imbalanced data classification. **Knowledge-Based Systems**, v. 242, p. 108295, abr. 2022.

REN, Z. *et al*. Adaptive cost-sensitive learning: Improving the convergence of intelligent diagnosis models under imbalanced data. **Knowledge-Based Systems**, v. 241, p. 108296, abr. 2022.

SELVA BIRUNDA, S.; KANNIGA DEVI, R. A Review on Word Embedding Techniques for Text Classification. *In*: REDDY, V. S. K.; VISWANATHAN, M.; ARUNA, S. (ed.). **Advances in Electrical, Communication and Computer Engineering**. Singapore: Springer, 2021. p. 267-281. (Lecture Notes on Data Engineering and Communications Technologies, v. 56). DOI: 10.1007/978-981-15-4635-4_24.

SIKDER, S. *et al*. Study of Machine Learning Techniques to Mitigate Fraudulent Transaction in Credit Cards. *In*: BHATTACHARYYA, S. *et al*. (ed.). **Innovative Computing and Communications**. Singapore: Springer Nature, 2022. p. 555-567. (Lecture Notes in Electrical Engineering, v. 876).

SINGH, A.; RANJAN, R. K.; TIWARI, A. Credit Card Fraud Detection under Extreme Imbalanced Data: A Comparative Study of Data-level Algorithms. **Journal of Experimental and Theoretical Artificial Intelligence**, v. 34, n. 4, p. 571–598, 2021.

TANG, X.; CHEN, L. Artificial bee colony optimization-based weighted extreme learning machine for imbalanced data learning. **Cluster Computing**, v. 22, n. S3, p. 6937–6952, 2019.

THORMUNDSSON, B. **Topic: Machine learning**. Statista, 3 mar. 2023. Disponível em: https://www.statista.com/topics/9583/machine-learning/. Acesso em: 11 jan. 2024.

TIAN, Y.; ZHANG, Y. A comprehensive survey on regularization strategies in machine learning. **Information Fusion**, Amsterdam, v. 80, p. 146-166, 2022. DOI: 10.1016/j.inffus.2021.11.005.

TRAN, T.; TRAN, L.; MAI, A. K-Segments Under Bagging approach: An experimental Study on Extremely Imbalanced Data Classification. *In*: INTERNATIONAL SYMPOSIUM ON COMMUNICATIONS AND INFORMATION TECHNOLOGIES (ISCIT), 19., 2019, [S. l.]. **Anais** [...]. [S. l.]: IEEE, 2019. p. 222-227. DOI: 10.1109/iscit.2019.8905145.

TRANFIELD, D.; DENYER, D.; SMART, P. Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. **British**

Journal of Management, v. 14, n. 3, p. 207–222, set. 2003.

TRISANTO, D. *et al*. Modified Focal Loss in Imbalanced XGBoost for Credit Card Fraud Detection. **International Journal of Intelligent Engineering and Systems**, v. 14, n. 4, p. 350–358, 2021.

VELANDIA-CARDENAS, C.; VIDAL, Y.; POZO, F. Wind Turbine Fault Detection Using Highly Imbalanced Real SCADA Data. **Energies**, v. 14, n. 6, p. 1728, 2021.

VERDONCK, T. *et al*. Special issue on feature engineering editorial. **Machine Learning**, New York, v. 110, n. 12, p. 3917-3928, 2021. DOI: 10.1007/s10994-021-06042-2.

WANG, D. *et al*. A Safe Zone SMOTE Oversampling Algorithm Used in Earthquake Prediction Based on Extreme Imbalanced Precursor Data. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 35, n. 13, out. 2021.

WANG, G.; CHEN, G.; CHU, Y. A new random subspace method incorporating sentiment and textual information for financial distress prediction. **Electronic Commerce Research and Applications**, v. 29, p. 30–49, maio 2018.

WANG, H.; ZHENG, H. True Positive Rate. *In*: DUBITZKY, W. *et al*. (ed.). **Encyclopedia of Systems Biology**. New York: Springer, 2013. p. 2302-2303.

WU, Z.; ZHANG, J.; HU, S. Review on Classification Algorithm and Evaluation System of Machine Learning. *In*: INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGY AND APPLICATIONS (ICICTA), 2020, [S. l.]. **Anais** [...]. [S. l.]: IEEE, 2020. p. 1-4. DOI: 10.1109/icicta51737.2020.00052.

XIA, H. *et al*. A multi-scale segmentation-to-classification network for tiny microaneurysm detection in fundus images. **Knowledge-Based Systems**, Amsterdam, v. 226, p. 107140, ago. 2021. DOI: 10.1016/j.knosys.2021.107140.

YUAN, Y. *et al*. Review of resampling techniques for the treatment of imbalanced industrial data classification in equipment condition monitoring. **Engineering Applications of Artificial Intelligence**, v. 126, p. 106911, nov. 2023.

ZHANG, X.; QIN, L. An Improved Extreme Learning Machine for Imbalanced Data Classification. **IEEE Access**, v. 10, p. 8634–8642, 2022.

ZHAO, S. *et al*. Mutation grey wolf elite PSO balanced XGBoost for radar emitter individual identification based on measured signals. **Measurement**, v. 159, p. 107777, jul. 2020.

ZHU, T. *et al*. Minority oversampling for imbalanced time series classification. **Knowledge-Based Systems**, v. 247, p. 108764, jul. 2022.

ZIĘBA, M. *et al*. Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. **Applied Soft Computing**, v. 14, p. 99–108, jan. 2014.

ANEXOS

Anexo A - Script em linguagem R para geração de dados sintéticos (exemplo)

```
system.time({


#StdOrder

id <- 0


#Define the levels of the experiment.

num_features <- 2 | 10

num_samples_multiplier <- 1000 | 1500

balancing <- 0.005 | 0.01

interaction <- FALSE | TRUE

threshold <- 0.5 | 0.8

error <- 0.2 | 0.8


#Generating the Features.

num_samples <- num_features * num_samples_multiplier

x <- matrix(runif(num_features * num_samples, min = -2, max = 2), ncol = num_features)

print(x)


#Functions without the error term.

function_2_no_error <- function(x) {

  return(0.5 + 0.2 * x[1] - 0.1 * x[2] )
```

```
}

function_interaction_2_no_error <- function(x) {

  return(0.5 + 0.2 * x[1] - 0.1 * x[2] + 0.1 * (x[1] * x[2]))

}

function_10_no_error <- function(x) {

  return(0.5 + 0.2 * x[1] + 0.1 * x[2] + 0.3 * x[3] + 0 * x[4] + 0.1 * x[5]

         - 0.2 * x[6] + 0.4 * x[7] + 0 * x[8] - 0.3 * x[9] + 0.2 * x[10])

}

function_interaction_10_no_error <- function(x) {

  return(0.5 + 0.2 * x[1] + 0.1 * x[2] + 0.3 * x[3] + 0 * x[4] + 0.1 * x[5]

         - 0.2 * x[6] + 0.4 * x[7] + 0 * x[8] - 0.3 * x[9] + 0.2 * x[10]

         + 0.2 * (x[1] * x[2])) - 0.1 * (x[3] * x[4] * x[5])

}



#Choice of the Function and Effect Size Calculation according to the levels {num_features and
interaction}.

  if(num_features == 2) {

    if (interaction == FALSE){

      y <- (apply(x, 1, function_2_no_error))

      print(efect <- (sd(y) * error))

      print ('f=2 and FALSE')

    } else{

      y <- (apply(x, 1, function_interaction_2_no_error))

      print(efect <- (sd(y) * error))

      print ('f=2 and TRUE')

    }
```

```
 }else{
   if (interaction == FALSE){
     y <- (apply(x, 1, function_10_no_error))
     print(efect <- (sd(y) * error))
     print ('f=10 and FALSE')
   }else{
     y <- (apply(x, 1, function_interaction_10_no_error))
     print(efect <- (sd(y) * error))
     print ('f=10 and TRUE')
   }
 }


#Linear Function:
#Two Variables: y = 0.5 + 0.2x1 - 0.1x2 + e
function_2 <- function(x) {
  return(0.5 + 0.2 * x[1] - 0.1 * x[2] + rnorm(1, mean = 0, sd = efect))
}


#Ten Variables: y = 0.5 + 0.2x1 − 0.1x2 + 0.3x3 + 0x4 + 0.1x5 − 0.2x6 +
#0.4x7 + 0x8 − 0.3x9 + 0.2x10 + e
function_10 <- function(x) {
  return(0.5 + 0.2 * x[1] + 0.1 * x[2] + 0.3 * x[3] + 0 * x[4] + 0.1 * x[5]
         - 0.2 * x[6] + 0.4 * x[7] + 0 * x[8] - 0.3 * x[9] + 0.2 * x[10] +
           rnorm(1, mean = 0, sd = efect))
}
```

```
linear_function <- ifelse(num_features == 2, function_2, function_10)


#Linear Function with interactions:
#Two Variables: y = 0.5 + 0.2x1 - 0.1x2 + 0.1x1x2 + e
function_interaction_2 <- function(x) {
  return(0.5 + 0.2 * x[1] - 0.1 * x[2] + 0.1 * (x[1] * x[2]) +
           rnorm(1, mean = 0, sd = efect))
}


#Ten Variables: y = 0.5 + 0.2x1 − 0.1x2 + 0.3x3 + 0x4 + 0.1x5 − 0.2x6 + 0.4x7 +
#0x8 − 0.3x9 + 0.2x10 + 0.2x1x2 -0.1x3x4x5+ e
function_interaction_10 <- function(x) {
  return(0.5 + 0.2 * x[1] + 0.1 * x[2] + 0.3 * x[3] + 0 * x[4] + 0.1 * x[5]
         - 0.2 * x[6] + 0.4 * x[7] + 0 * x[8] - 0.3 * x[9] + 0.2 * x[10]
         + 0.2 * (x[1] * x[2]) - 0.1 * (x[3] * x[4] * x[5])
         + rnorm(1, mean = 0, sd = efect))
}


linear_function_interaction <-
  ifelse(num_features == 2, function_interaction_2, function_interaction_10)


#Checking Function.
print(ifelse(interaction == FALSE, linear_function, linear_function_interaction))
```

```r
#Generating y.

y_raw <- apply(x, 1, ifelse(interaction == FALSE, linear_function, linear_function_interaction))


#Round Y values to 1 if they are greater than or equal to the threshold, otherwise round to 0.

y <- ifelse(y_raw >= threshold, 1, 0)

print(y)

table(y)


#Calculating the quantity of y = 1 for the imbalance.

minority_class <- num_samples * balancing

print (minority_class)


#Count the number of zeros and ones in vector Y.

num_zeros <- sum(y == 0)

num_uns <- sum(y == 1)


#Calculate the minimum number of observations to remove.

obs_remover <- num_uns - minority_class


#Randomly remove observations from the dominant class until you reach the desired balancing.

if (num_uns > minority_class) {

  indices_remover <- sample(which(y == 1), obs_remover)

}


#Remove selected observations.
```

```
y_unbalancing <- y[-indices_remover]

table(y_unbalancing)

x_unbalancing <- x[-indices_remover, ]


#Joining X and y.

matriz <- data.frame(x_unbalancing, y = y_unbalancing)


#Create copies of Y = 0 to fill in the removed observations and return to the initial num_samples.

#Check if the number of elements to be sampled (size) is greater than the number of elements available in the vector.

set_replace = FALSE

if (length(matriz$y==0) < obs_remover) {

  set_replace = TRUE

}

indices_selecionados <- sample(which(matriz$y == 0), replace = set_replace)

matriz_copia <- matriz[indices_selecionados, ]


syntetic_base <- rbind(matriz, matriz_copia)


table(syntetic_base$y)


#Checking extremely imbalance.

balancing_y = sum (syntetic_base$y == 1) / num_samples

print(ifelse(balancing_y == balancing, 'Extremely Imbalanced', 'Fail'))


#Exporting to CSV.
```

```
path <- paste0('/home/leandro/ExperimentosTese/bases/synthetic_base_id_', id, '.csv')

write.csv(syntetic_base ,file =  path, row.names = FALSE)

})
```

Anexo B - Exemplo de um Script em Python para executar experimentos de modelos de classificação[10]

```
# -*- coding: utf-8 -*-

"""#2.ipynb


Automatically generated by Colab.


Original file is located at

    https://colab.research.google.com/drive/1jeoDmylSc7Mme5IkjeUd6qUUqVB-z6HP


#**Análise e importação de dados**

"""


import pandas as pd


base = pd.read_csv('Downloads/synthetic_base_id_2.csv')

base


"""##**Analisando o desbalanceamento**"""


import numpy as np

np.unique(base['y'], return_counts= True)
```

[10] Os scripts para todas as 32 execuções podem ser visualizados em: <https://drive.google.com/drive/folders/1cHtreEGafgh_wCmA2LdT2nFQP3ULoK1S?usp=sharing>

```
import seaborn as sns

sns.countplot(x = base['y'])


"""##**Extração dos vetores**"""


X_base = base.iloc[:, 0:10].values

X_base.shape, X_base


Y_base = base.iloc[:, 10].values

Y_base.shape, Y_base


"""###**Analisando a separação entre classes**"""


# Aplicar PCA para reduzir para 2 dimensões e gerar Scatter Plot

from sklearn.decomposition import PCA


pca = PCA(n_components=2)

X_pca = pca.fit_transform(X_base)


import matplotlib.pyplot as plt


# Scatter plot

for i in range(len(X_pca)):

    if Y_base[i] == 0:

        plt.scatter(X_pca[i][0], X_pca[i][1], color='red')
```

```
    else:
        plt.scatter(X_pca[i][0], X_pca[i][1], color='blue')

plt.xlabel('X1')
plt.ylabel('X2')
plt.title('Scatter Plot of Binary Classification')
plt.show()

"""##**SMOTE**"""

from imblearn.over_sampling import SMOTE

smote = SMOTE(sampling_strategy='minority')
X_smote, Y_smote = smote.fit_resample(X_base, Y_base)

print("Contagem de classes após o SMOTE:")
unique, counts = np.unique(Y_smote, return_counts=True)
print(dict(zip(unique, counts)))

X_smote.shape, X_smote, Y_smote.shape, Y_smote

"""##**Tuning de parâmetros com Grid Search**

Com base no valor de F1-Score (VELANDIA-CARDENAS; VIDAL e POZO, 2021)
"""
```

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import make_scorer, f1_score

scorer = make_scorer(f1_score)

parametros = {'criterion': ['gini','entropy'],
              'n_estimators': [10, 50, 100],
              'max_depth': [None, 10, 20],
              'min_samples_split': [2, 5, 10],
              'min_samples_leaf': [1, 2, 4]}

grid_search = GridSearchCV(estimator = RandomForestClassifier(), param_grid = parametros,
scoring=scorer)
grid_search.fit(X_smote, Y_smote)
melhores_parametros = grid_search.best_params_
melhor_resultado = grid_search.best_score_
print(melhores_parametros)
print(melhor_resultado)

rf = RandomForestClassifier(criterion=melhores_parametros['criterion'],
                            min_samples_leaf=melhores_parametros['min_samples_leaf'],
                            min_samples_split=melhores_parametros['min_samples_split'],
                            n_estimators=melhores_parametros['n_estimators'],
                            max_depth=melhores_parametros['max_depth'])
```

```
rf.criterion, rf.max_depth, rf.min_samples_leaf, rf.min_samples_split, rf.n_estimators

"""##**RF Classificador e Validação Cruzada com K-fold**"""

from sklearn.model_selection import cross_val_score, KFold

"""###Accuracy (**ACC**)

"""

resultado_acc = []
for i in range(30):
  kfold = KFold(n_splits=10, shuffle=True, random_state=i)

  scores = cross_val_score(rf, X_smote, Y_smote, cv=kfold)
  print (scores)
  print (scores.mean())
  resultado_acc.append(scores.mean())

"""###Precision (**PPV**)"""

resultado_ppv = []
for i in range(30):
  kfold = KFold(n_splits=10, shuffle=True, random_state=i)
```

```
 scores = cross_val_score(rf, X_smote, Y_smote, cv=kfold, scoring='precision')
 print (scores)
 print (scores.mean())
 resultado_ppv.append(scores.mean())

"""###Recall (**TPR**)"""

resultado_tpr = []
for i in range(30):
 kfold = KFold(n_splits=10, shuffle=True, random_state=i)

 scores = cross_val_score(rf, X_smote, Y_smote, cv=kfold, scoring='recall')
 print (scores)
 print (scores.mean())
 resultado_tpr.append(scores.mean())

"""###F1-Score (**F1**)"""

resultado_fs = []
for i in range(30):
 kfold = KFold(n_splits=10, shuffle=True, random_state=i)

 scores = cross_val_score(rf, X_smote, Y_smote, cv=kfold, scoring='f1')
 print (scores)
```

```
  print (scores.mean())

  resultado_fs.append(scores.mean())


"""###AUC-ROC (**ROC**)"""


resultado_roc = []

for i in range(30):

  kfold = KFold(n_splits=10, shuffle=True, random_state=i)


  scores = cross_val_score(rf, X_smote, Y_smote, cv=kfold, scoring='roc_auc')

  print (scores)

  print (scores.mean())

  resultado_roc.append(scores.mean())


"""###G-Mean (**G-M**)"""


from sklearn.metrics import make_scorer

from imblearn.metrics import geometric_mean_score


gmean_scorer = make_scorer(geometric_mean_score)


resultado_gmean = []

for i in range(30):

  kfold = KFold(n_splits=10, shuffle=True, random_state=i)
```

```
 scores = cross_val_score(rf, X_smote, Y_smote, cv=kfold, scoring=gmean_scorer)
 print (scores)
 print (scores.mean())
 resultado_gmean.append(scores.mean())


"""##**Resultados**"""

resultados = pd.DataFrame({'TPR':resultado_tpr, 'ROC':resultado_roc, 'F1':resultado_fs,
                           'PPV':resultado_ppv, 'ACC':resultado_acc, 'G-Mean':resultado_gmean})

"""###**Geral**"""

resultados.describe()

"""###**Coeficiente de Variação**

"""

(resultados.std() / resultados.mean() ) * 100
```

Anexo C - Certificado de Registro de Software (Script em linguagem R para geração de dados sintéticos)

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512025001786-3**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1° de janeiro subsequente à data de 24/04/2025, em conformidade com o §2°, art. 2° da Lei 9.609, de 19 de Fevereiro de 1998.

Título: SCRIPT GERADOR DE BASES SINTÉTICAS PARA CLASSIFICAÇÃO BINÁRIA EXTREMAMENTE DESBALANCEADA

Data de publicação: 24/04/2025

Data de criação: 31/03/2025

Titular(es): UNIFEI - UNIVERSIDADE FEDERAL DE ITAJUBÁ

Autor(es): PEDRO PAULO BALESTRASSI; LEANDRO DUARTE PEREIRA; FABRÍCIO ALVES DE ALMEIDA

Linguagem: R

Campo de aplicação: IF-07

Tipo de programa: IA-01

Algoritmo hash: SHA-512

Resumo digital hash:
4ba131c8bbd43f4adea07877bb3d1321869eb7fe2c784eb9e9b85b614163e598ef621825a5efbd2688a7842b09d521e317caa14e2bc2673b79ecd951a16d1f7e

Expedido em: 13/05/2025

Aprovado por:
Carlos Alexandre Fernandes Silva
Chefe da DIPTO

Anexo D - Artigo publicado em periódico internacional[11]

ELSEVIER

Information and Software Technology

Volume 187, November 2025, 107867

Review Article

Efficient approaches for binary classification in extremely imbalanced databases: A systematic literature review

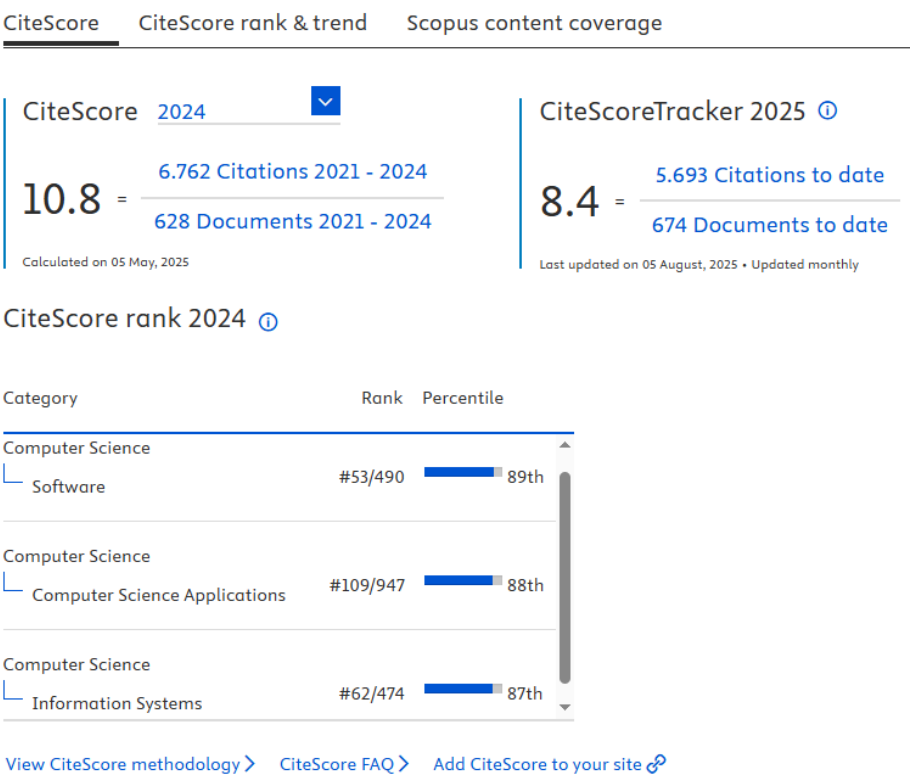
Leandro Duarte Pereira [a], Fabrício Alves de Almeida [a], Farid Melgani [b], Pedro Paulo Balestrassi [a]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.infsof.2025.107867 ↗ Get rights and content ↗

Full text access



CiteScore CiteScore rank & trend Scopus content coverage

CiteScore 2024

10.8 = 6.762 Citations 2021 - 2024 / 628 Documents 2021 - 2024

Calculated on 05 May, 2025

CiteScoreTracker 2025

8.4 = 5.693 Citations to date / 674 Documents to date

Last updated on 05 August, 2025 • Updated monthly

CiteScore rank 2024

Category	Rank	Percentile
Computer Science: Software	#53/490	89th
Computer Science: Computer Science Applications	#109/947	88th
Computer Science: Information Systems	#62/474	87th

View CiteScore methodology CiteScore FAQ Add CiteScore to your site

[11] Acesso completo em: <https://doi.org/10.1016/j.infsof.2025.107867>

Anexo E - Artigo publicado em periódicos nacionais [Qualis A3] [12]

Conceitos e Métodos de Inteligência Artificial para Análise de Dados

Artificial intelligence concepts and methods for data analysis

Gabriel Mendes Cirac[1], Tiago Cavalca[2], Luiz Fernando Nunes[3], Leandro Duarte Pereira[4], Valdinei de Paula Rodrigues[5]

1 – Doutor em Engenharia Elétrica – Unifei. E-mail.: gabriel_cirac@yahoo.com.br
2 – Bacharel em Sistemas de Informação pela FEPI. E-mail.:tiagocavalca47@gmail.com
3 – Mestre em Ciência e Tecnologia da Informação. Professor da FEPI. E-mail.: luiz.nunes@fepi.br
4 – Mestre em Engenharia de Produção e Professor da FEPI. E-mail.: leandro.pereira@fepi.br
5 – Mestre em Engenharia Eletrônica e Computação. Professor da FEPI. E-mail.: prof.valdinei.rodrigues@gmail.com

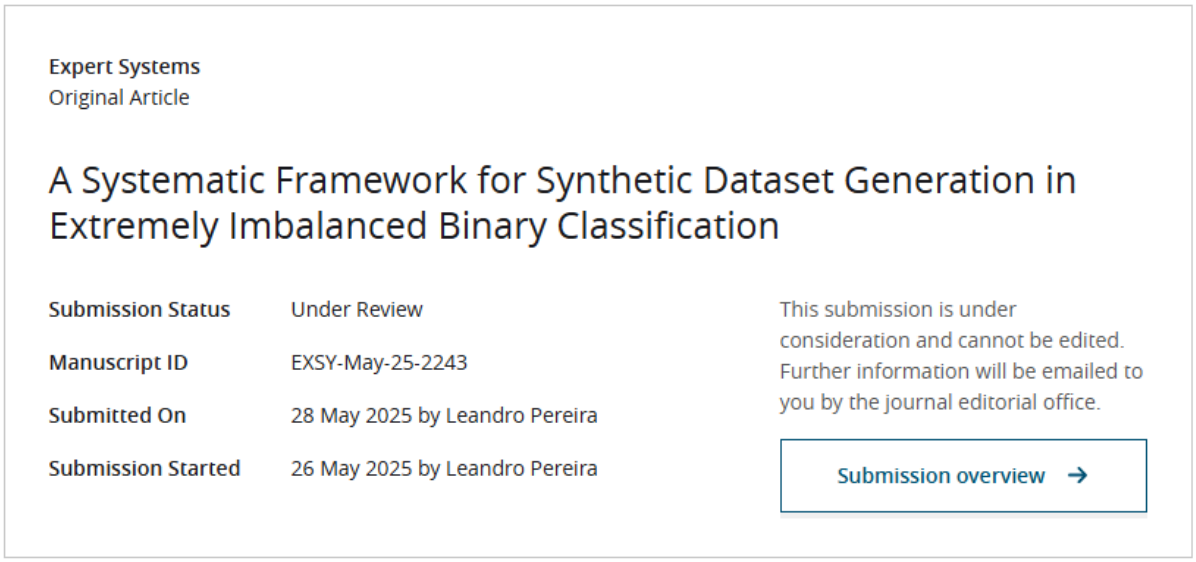
Recebido em: 25/09/2023 Revisado em: 17/10/2024 Aprovado em: 16/12/2024

Resumo: Inteligência Artificial (IA) é um conceito antigo, com debates especulativos datando de dois milênios, quando filósofos sonhavam em replicar a lógica humana por meio de dispositivos. Na origem, visavam entender processos de aprendizagem, memória, visão e raciocínio, visando a mecanização da inteligência. Com a evolução tecnológica, a IA ressurgiu como tópico relevante, introduzido sistematicamente na década de 1950 pelo artigo "Computing Machinery And Intelligence", prenunciando o potencial do aprendizado de máquina. Os avanços e diversificação da IA têm facilitado a vida humana, aplicando-se desde dispositivos eletrônicos, como Alexa, até análises de marketing e financeiras, auxiliando na tomada de decisões e automatização de processos. Este trabalho revisa a literatura de IA e realiza um estudo de caso com algoritmos de Classificação (Árvore de decisão, SVM e SGDClassifier), Agrupamento (K-Means), método Average Silhouette e decomposição PCA, dividindo os dados para treinamento, com foco na análise e exploração de dados de consumo. Os resultados são apresentados através de tabelas e gráficos comparativos, onde é possível verificar padrões de compras e perfis de consumo.
Palavras-chave: Inteligencia Artificial. Análise de Dados. Python. Aprendizado de Máquina.

Abstract: Artificial Intelligence (AI) is an ancient concept, with speculative debates dating back two millennia when philosophers dreamed of replicating human logic through devices. Originally, they aimed to understand processes of learning, memory, vision, and reasoning, seeking the mechanization of intelligence. With technological evolution, AI re-emerged as a relevant topic, systematically introduced in the 1950s by the article "Computing Machinery and Intelligence," foreshadowing the potential of machine learning. The advances and diversification of AI have facilitated human life, being applied from electronic devices like Alexa to marketing and financial analyses, assisting in decision-making and process automation. This work reviews the AI literature and conducts a case study with Classification algorithms (Decision Tree, SVM, and SGDClassifier), Clustering (K-Means), the Average Silhouette method, and PCA decomposition, splitting the data for training, focusing on the analysis and exploration of consumption data. The results are presented through comparative tables and graphs, where it is possible to identify purchasing patterns and consumption profiles.
Keywords: Artificial Intelligence, Data Analysis. Python. Machine Learning.

[12] Acesso completo em: <http://revista.fepi.br/revista/index.php/revista/article/download/902/pdf_209>

Anexo F - Submissões em periódico internacional



Expert Systems
Original Article

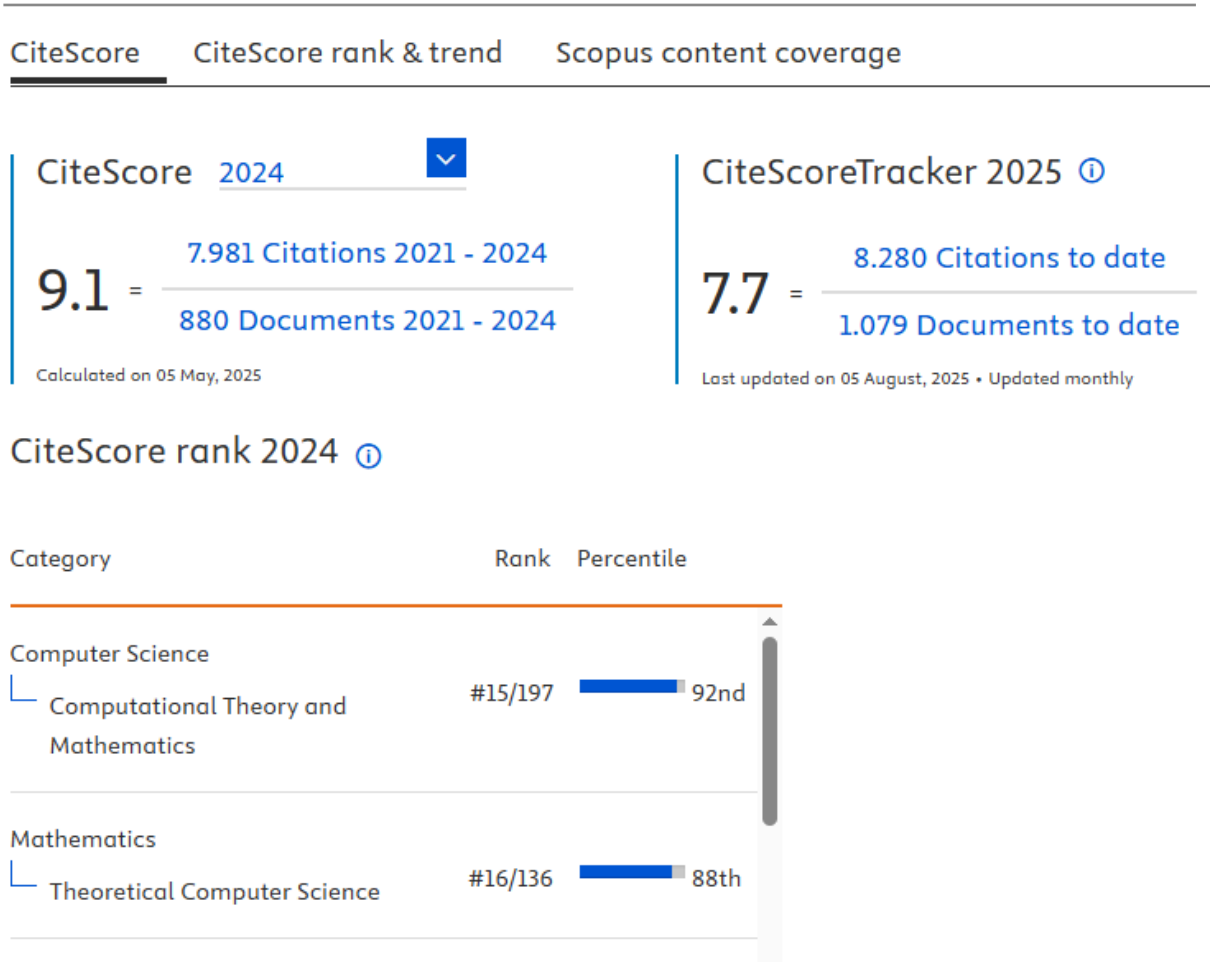
A Systematic Framework for Synthetic Dataset Generation in Extremely Imbalanced Binary Classification

Submission Status	Under Review
Manuscript ID	EXSY-May-25-2243
Submitted On	28 May 2025 by Leandro Pereira
Submission Started	26 May 2025 by Leandro Pereira

This submission is under consideration and cannot be edited. Further information will be emailed to you by the journal editorial office.

Submission overview →

Status atualizado em 18/08/2025.



CiteScore | CiteScore rank & trend | Scopus content coverage

CiteScore 2024

9.1 = 7.981 Citations 2021 - 2024 / 880 Documents 2021 - 2024

Calculated on 05 May, 2025

CiteScoreTracker 2025

7.7 = 8.280 Citations to date / 1.079 Documents to date

Last updated on 05 August, 2025 • Updated monthly

CiteScore rank 2024

Category	Rank	Percentile
Computer Science Computational Theory and Mathematics	#15/197	92nd
Mathematics Theoretical Computer Science	#16/136	88th