Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Predição de falência utilizando dados sequenciais não estacionários em uma abordagem de fluxo de dados

Rubens Marques Chaves

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientador
Prof. Dr. Luís Paulo Faina Garcia

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Predição de falência utilizando dados sequenciais não estacionários em uma abordagem de fluxo de dados

Rubens Marques Chaves

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof. Dr. Luís Paulo Faina Garcia (Orientador)
CIC/UnB

Prof. Dr. Ricardo Cerri
Membro externo - DC/UFSCar

Prof. Dr. Célia Ghedini Ralha
Membro interno - CIC/UnB

Prof. Dr. Thiago de Paulo Faleiros
Suplente - CIC/UnB

Prof. Dr. Ricardo Pezzuol Jacobi
Coordenador do Programa de Pós-graduação em Informática

Brasília, 20 de julho de 2023

Dedicatória

> O valor de todo o conhecimento está no seu vínculo com as nossas necessidades, aspirações e ações; de outra forma, o conhecimento torna-se um simples lastro de memória, capaz apenas – como um navio que navega com demasiado peso – de diminuir a oscilação da vida quotidiana.
>
> *V. O. Kliutchevski (1841-1911)*

Dedico este trabalho, primeiramente, à minha esposa que me apoiou e incentivou para buscar aperfeiçoamento profissional por meio do mestrado, e a todo povo desta nação, que contribuiu com este estudo através da licença capacitação concedida pelo Banco Central do Brasil. Aos que antes de mim trilharam o caminho da ciência e pavimentaram muitas das estradas que precisei percorrer para produzir este estudo e, por fim, aos que virão, espero que façam bom proveito do conteúdo compilado neste trabalho e que lhes ajude, como muitas vezes fui ajudado por trabalhos anteriores.

Agradecimentos

Nesse período a ajuda de pessoas a minha volta foi fundamental para conduzir esse estudo. Primeiramente ao meu orientador, Dr. Luís Paulo, que com paciência e ótimas dicas tem me ajudado a elevar o nível desta pesquisa para produzir algo relevante ao meio acadêmico. A Maria, mãe de minha esposa, por todo suporte dado em nossa casa. A minha esposa, Sabrina, que muitas vezes precisou fazer as vezes de pai, quando estive ausente devido a pequisa, especialmente junto aos nossas filhas Maria Helena, Angelina, Lara, Valentina e nosso filho Benjamim. A elas e ele que me ajudaram a relaxar nos momentos de mais estresse, sempre me ajudando a relembrar como é bom ser criança e levar as coisas com mais leveza. Por fim, a Deus que por meio desses tem demonstrado o seu amor por mim e me ajudado a encontrar a paz necessária para produzir ciência.

Resumo

As previsões de falência corporativa são cruciais para empresas, investidores e autoridades. No entanto, muitos estudos de previsão de falências se basearam em modelos estacionários e ignoraram desafios importantes, como a dependência temporal, desvio de conceito e desequilíbrio de dados. O objetivo deste estudo é propor métodos para enfrentar esses desafios e utilizar dados das demonstrações financeiras trimestrais obtidas de empresas que se reportam à Comissão de Valores Mobiliários (CVM). O conjunto de dados abrange dez anos (2011 a 2020) e inclui 905 empresas diferentes com 23.834 registros, cada um contendo 84 indicadores. Enquanto a maioria das empresas da amostra não apresenta dificuldades financeiras, 651 registros são de empresas com dificuldades financeiras. Para lidar com o desvio de conceito, o experimento empírico pré-processa os dados usando uma janela deslizante, histórico e mecanismo de esquecimento para lidar com desvio de conceito. Além disso, diversas técnicas de balanceamento são empregadas para lidar com o desequilíbrio dos dados. Modelos de aprendizado supervisionado, entre eles Regressão Logística (RL), Máquina de Vetores de Suporte (*Support Vector Machine* - SVM), *Random Forest* (RF), Árvore de Decisão (AD), *Extreme Gradient Boosting* (XGBoost) e *Categorial Boosting* (CatBoost), são avaliados para identificar o modelo de melhor desempenho. Dadas as características do problema, particularmente o desequilíbrio de dados, o desempenho do modelo é medido usando métricas como Precisão, Sensibilidade, F_1-Score, G_{mean}, *Area Under the Curve - ROC* (AUC-ROC) and *Area Under the Curve - Precision and Sensitivity* (AUC-PS). Os modelos de melhor desempenho (RF, XGBoost e CatBoost) indicam que a abordagem adotada gera ganho de desempenho.

Palavras-chave: Fluxo de Dados, Desbalanceamento, Falência, Dificuldade Financeira

Abstract

Corporate bankruptcy predictions are crucial for companies, investors, and authorities. However, many bankruptcy prediction studies have relied on stationary models and overlooked important challenges such as data non-stationarity, concept drift, and data imbalance. This study proposes methods to address these challenges and utilize quarterly financial statement data obtained from companies reporting to the Securities and Exchange Commission of Brazil (CVM). The dataset covers ten years (2011 to 2020) and includes 905 different companies with 23,834 records, each containing 84 indicators. While most of the companies in the sample do not present financial difficulties, 651 records are from companies facing financial challenges. To address the concept drift, the empirical experiment preprocesses the data using a sliding window, historical data, and a forgetting mechanism to handle concept drift. Furthermore, various balancing techniques are employed to tackle the data imbalance. Supervised learning models, such as Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Extreme Gradient Boosting (XGBoost), and Categorial Boosting (CatBoost), are evaluated to identify the best-performing model. Given the characteristics of the problem, particularly the data imbalance, model performance is measured using Precision, Sensitivity, F_1-Score, G_{mean}, Area Under the Curve - ROC (AUC-ROC) and Area Under the Curve - Precision and Sensitivity (AUC-PS). The best-performing models (RF, XGBoost and CatBoost) indicate that the adopted approach generates performance gain.

Keywords: Data Stream, Imbalance data, Bankruptcy, Financial distress

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AD Árvore de Decisão.

ADAMS Advanced Data mining And Machine learning System.

ADASYN Adaptive Synthetic.

ADL Análise Discriminante Linear.

ADM Análise de Discriminante Múltipla.

ADWIN Adaptive Windowing.

AG Algoritmo genético.

AM Aprendizado de Máquina.

ANS Adaptive Neighbor SMOTE.

AUC-PS Area Under the Curve - Precision and Sensitivity.

AUC-ROC Area Under the Curve - ROC.

AWSMOTE Adaptive-Weighting SMOTE.

B-SMOTE Borderline-SMOTE.

B3 Bolsa de Valores do Brasil.

BCB Banco Central do Brasil.

BPA Balanço Patrimonial de Ativos.

BPP Balanço Patrimonial de Passivos.

CA Capital Acumulado.

CAPES Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

CatBoost Categorial Boosting.

CM Competence Model.

CUSUM Cumulative Sum.

CVM Comissão de Valores Mobiliários.

DCF Dynamic Clustering Forest.

DDE Drift Detection Ensemble.

DDM Drift Detection Method.

DDM-FP-M Drift Detection Method with False Positive rate for Multi-label classification.

DELM Dynamic Extreme Learning Machine.

DF Dificuldade Financeira.

DFC Demonstração de Fluxo de Caixa.

DFT Discrete Fourier Transform.

DMDDM Diversity Measure as a new Drift Detection Method.

DRE Demonstração de Resultado.

ECDD Ensemble Classifiers with Drift Detection.

EDDM Early Drift Detection Method.

EDE Equal Density Estimation.

EDTC Ensemble Decision Trees for Concept drift.

EUA Estados Unidos da América.

FHDDM Fast Hoeffding Drift Detection Method.

FIWCD Fourier Inspired Windows for Concept Drift.

FMI Fundo Monetário Internacional.

FN Falso Negativo.

FP Falso Positivo.

FP-ELM Forgetting Parameters Extreme Learning Machine.

FRE Formulário de Referência.

FTDD Fisher Test Drift Detector.

HDDM Hoeffding Drift Detection Method.

HHT-AG Hierarchical Hypothesis Testing with Attribute-wise "Goodness-of-fit".

HHT-CU Hierarchical Hypothesis Testing with Classification Uncertainty.

HLFR Hierarchical Linear Four Rates.

IA Inteligência Artificial.

ITR Formulário de Informações Trimestrais.

KNN K-Nearest Neighbor.

LDD-DSDA Local Drift Degree based Density Synchronized Drift Adaptation.

LLDD Learning with Local Drift Detection.

LMT Logistic Model Tree.

LSDD Least Squares Density Difference.

LSDD-CDT Least Squares Density Difference based on Change Detection Test.

MD3 Margin Density Drift Detection.

MOA Massive Online Analysis.

N Total de Negativos Reais.

NN-DVI Nearest Neighbor based on Density Variation Identification method.

NP Total de Negativos Preditos.

OCDD One Class Drift Detector.

OS-ELM Online Sequential Extreme Learning Machine.

P Total de Positivos Reais.

PCA-CD Principal Component Analysis based on Change Detection.

PHT Page-Hinkley Test.

PP Total de Positivos Preditos.

PS Curva de Precisão e Sensibilidade.

RBC Raciocínio Baseado em Casos.

RDDM Reactive Drift Detection Method.

Re-DBSCAN Revising Density-based Spatial Clustering of Applications with Noise.

REA Recursive Ensemble Approach.

RF Random Forest.

RL Regressão Logística.

RNA Redes Neurais Artificiais.

RNC Rede Neural Convolucional.

ROA Return on Assets.

ROC Receiver Operating Curve.

ROE Return on Equity.

SAMOA Scalable Advanced Massive Online Analysis.

SCD Statistical Change Detection.

SDF Sem Dificuldade Financeira.

SFN Sistema Financeiro Nacional.

SMOTE Synthetic Minority Over-Sampling Technique.

SMOTE-ENN SMOTE with Edited Nearest Neighbor.

SMOTE-Tomek SMOTE with Tomek Links.

SNDC Self-adaption Neighborhood Density Clustering method.

STDS Self-Training Data Streams.

STEPD Statistical Test of Equal Proportions.

SVM Support Vector Machine.

SVM-SMOTE Support Vector Machine-SMOTE.

TFP Taxa de Falso Positivo.

TSMSD-EWMA Two-Stage Multivariate Shift-Detection test based on EWMA.

TVN Taxa de Verdadeiro Negativo.

TVP Taxa de Verdadeiro Positivo.

UCI University of California, Irvine.

VN Verdadeiro Negativo.

VP Verdadeiro Positivo.

VW Vowpal Wabbit.

WSTD Wilcoxonr rank Sum Test Drift detector.

XGBoost Extreme Gradient Boosting.

Capítulo 1

Introdução

Atualmente, estamos inseridos em um ambiente cercado de dispositivos capazes de coletar informações do usuário e/ou do ambiente, de maneira frequente, praticamente online, ou tendo que lidar com informações de múltiplas fontes, algumas vezes extrapolando as fronteiras do país. Assim, dia após dia, cada vez mais rápido, estamos nos tornando digitais e deixando rastros de dados durante nossa existência que podem ser utilizados em algoritmos de Inteligência Artificial (IA) [1, 2].

As consequências do fracasso financeiro são enormes para credores, gerentes, acionistas, investidores, funcionários e até mesmo para a economia de um país, diante da forma como as sociedades hodiernas estão estruturadas e os mercados econômicos funcionam. É por isso que, desde a década de 1960 [3], prever a falência corporativa tornou-se uma preocupação significativa para os vários *stakeholders* das empresas. A predição de falência leva a benefícios, como redução de custos na análise de crédito, melhor monitoramento e aumento da taxa de cobrança de dívidas [4]. Por isso, as empresas no mercado de capitais devem fornecer aos interessados (*i.e.* investidores e instituições financeiras fornecedoras de empréstimos) informações sobre o negócio. As informações são usadas para avaliar a empresa como uma possibilidade de investimento ou um possível cliente para um empréstimo. Dessa forma, é possível evitar prejuízos futuros em casos de Dificuldade Financeira (DF) ou falência. Essas informações devem ser fornecidas periodicamente à Comissão de Valores Mobiliários (CVM) e podem sofrer impactos de diferentes tipos como questões climáticas, culturais, econômicas, sanitárias, entre outros [5]. Esse é um cenário típico de fluxo de dados e precisa de um tratamento adequado [6].

Predizer falências tornou-se uma área de grande interesse e está ganhando importância atualmente [7]. Hoje, a questão não é se devemos usar modelos preditivos de falências, mas como aumentar o desempenho dos modelos existentes. Existem duas principais vertentes de modelos que têm sido usados para prever falências: (*i*) modelos estatísticos e (*ii*) modelos de IA [4]. Desde a década de 60, estudos têm sido direcionados à questão de

predição de DF, a maioria utilizando ferramentas estatísticas [3, 8, 9], com destaque para as técnicas de Análise de Discriminante Múltipla (ADM) e modelos de Regressão Logística (RL) [9]. Entretanto, os modelos estatísticos têm algumas limitações, como linearidade, normalidade e independência entre as variáveis [4].

Desde a década de 1990, a academia tem buscado alternativas aos métodos estatísticos, assim, abordagens de IA tem sido utilizadas, como por exemplo o Aprendizado de Máquina (AM) [10]. Barboza, F. *et al.* (2017) [7], verificaram que modelos preditivos utilizando AM apresentam resultados melhores que os modelos estatísticos. É difícil atribuir esse aumento de popularidade a um único fator. Nesse período houve um grande aumento na quantidade de dados gerados e armazenados, requerendo uma forma diferente de processa-los. Dessa forma, popularizou-se o termo *Big Data* [11, 12] que vem estimulando a utilização de diferentes ferramentas para o processamento dos dados, entre elas estão algumas técnicas de AM [13]. Além disso, diversas bibliotecas e ferramentas de AM vêem sendo desenvolvidas e disponibilizadas, como: Weka [14], scikit-learn [15], Tensorflow [16] e Pytorch [17]; que possuem vários algoritmos e funcionalidades já implementadas, prontas para o uso e sem a necessidade de configurações complexas.

A quantidade de dados gerada nos dias de hoje é de difícil armazenagem e processamento. Tomemos como exemplo o ano de 2012, quando segundo estimativas relatadas em *IDC Survey* [12] foram gerados mais de 2,8 zetabytes (2,8 trilhões de gigabytes). O processamento de quantidade tão expressiva de dados representa um desafio até mesmo para processos em *batch* [18]. Entretanto, essa não é a única dificuldade, pois novos dados têm sido gerados constantemente de maneira semelhante a um riacho com fluxo de água contínuo. É o que chamamos fluxo de dados (*data stream*). Assim, podemos dizer que, qualquer sequência de dados com um carimbo de data/hora (*timestamp*) é conhecida como fluxo de dados, que contém um grande volume de dados com velocidades variadas. Os dados digitais gerados a partir de diferentes fontes têm as seguintes características [19]:

- Os fluxos de dados são perenes (tamanho ilimitado).
- Os dados são recebidos ou coletados em taxa e velocidade variáveis.
- Com o passar do tempo os dados evoluem.
- O processamento dos dados tem restrições de memória.

É comum, seja de maneira explícita ou implícita, considerar que o processo de geração de fluxo de dados é estacionário, isto é, os dados são extraídos de uma distribuição de probabilidade fixa, embora desconhecida [20]. Todavia, no mundo real, as aplicações precisam lidar com mudanças rápidas, como: novos produtos, mercados e comportamentos do cliente; resultando em um fluxo de dados não estacionário (em evolução) [6]. A não

estacionariedade pode ser devido, por exemplo, a efeitos de sazonalidade ou periodicidade, mudanças nos hábitos ou preferências dos usuários, falhas de *hardware* ou *software* que afetam sistemas ciberfísicos (*Ciber-Physics Systems*), desvios térmicos ou efeitos de envelhecimento em sensores [20]. Esse comportamento, inerente aos fluxos de dados reais, é conhecido por desvio de conceito (*concept drift*) [21, 22] e precisa ser tratado para evitar a degradação do modelo. Quando o impacto dessas mudanças é pequeno, o modelo não se comporta de maneira ideal e, nos piores casos, o modelo apresenta erros extremamente grosseiros [20].

O desvio de conceito pode ser visto em diferentes domínios onde as previsões são ordenadas por tempo. Por exemplo, a previsão de condições atmosféricas tem três atributos: temperatura, umidade e pressão. A estação do ano (primavera, verão, outono, inverno) pode ser um conceito no fluxo de dados climáticos que afeta a temperatura (ou seja, não é explicitamente especificado nos dados de temperatura), mas pode influenciar os dados climáticos. Assim, informações ocultas (*i.e.* a estação do ano) podem causar desvio de conceito [19]. Outro exemplo é o comportamento de compra dos clientes ao longo do tempo que pode influenciar a força da economia de um país. O padrão de compra dos clientes pode mudar com o tempo, dependendo do dia da semana, disponibilidade de alternativas, taxa de inflação, entre outras razões [23]. Dessa forma, os dados coletados em um determinado período mostram a mudança na relação do comportamento de compra do cliente com a força da economia [19]. Outro caso seria o impacto da inflação ou de uma pandemia sobre o resultado financeiro de uma ou mais empresas, e ainda o impacto da sazonalidade nas vendas de um produto. Essas mudanças nos dados também podem alterar o balanceamento das classes, isto é, a proporção de determinadas classes presentes na amostra, o que também pode impactar no desempenho do modelo [19].

Gama, J. (2014) [18] define o desvio de conceito observando o comportamento dos dados em momentos distintos. Assim, quando a distribuição conjunta das variáveis de entrada e das classes é alterada de um momento para o outro dizemos que houve o desvio de conceito. Em outros termos, essa mudança nos dados pode ser caracterizada pela mudança nos componente da relação, sendo: (*i*) mudança na distribuição das classes; (*ii*) mudança na probabilidade condicional; e (*iii*) mudança na probabilidade posterior de classes. A primeira medida leva a um problema de desequilíbrio de classe e, por conseguinte, a degradação do desempenho do modelo [19].

O problema de degradação de desempenho do modelo preditivo, decorrente da mudança na distribuição de classes, é o objeto deste estudo, que tenta minimizá-lo. O desequilíbrio de classe ocorre quando instâncias na classe alvo (*i.e.* classe positiva ou minoritária) são fortemente sub-representadas em comparação com outra classe (*i.e.* classe negativa ou majoritária). Classificadores tradicionais tendem a ter viés para classe ma-

joritária e não podem acomodar o desequilíbrio de classe [24]. Esse problema se agrava quando a classe minoritária contém informações mais relevantes, como é o caso da predição de DF. Por exemplo, consideremos um conjunto de empresas interessadas em obter financiamento. Caso empresas saudáveis sejam classificadas como não aptas (falso positivo), elas não conseguirão receber o recurso e o problema estará restrito a elas. Por outro lado, quando empresas em DF são classificadas como aptas (falso negativo), elas receberão um recurso que não será restituído. O não pagamento dos empréstimos pode colocar a instituição financeira em situação de DF, podendo evoluir para uma falência e, por conseguinte, afetar todo o Sistema Financeiro Nacional (SFN) em uma reação em cadeia [25].

Normalmente, situações reais apresentam desequilíbrio de classes o que tem atraido a atenção da academia e do mercado, pois ocasiona uma dificuldade de avaliação dos modelos devido ao viés à classe majoritária [26]. Por isso, diversas abordagens foram propostos para minimizar o impacto desse problema, algumas por meio do método de balanceamento em algoritmo, outros através do pré-processamento dos dados ou pela junção de várias técnicas, conhecido por aprendizado por comitê. Entre esses métodos, as técnicas de reamostragem têm sido utilizadas com mais frequência [27].

Este estudo propõe uma solução para minimizar a degradação do modelo preditivo no tempo em decorrência do desbalanceamento de classes e desvio de conceito. Dessa maneira, contribui para preencher uma lacuna no meio acadêmico, a questão da falência ou DF em fluxo de dados desbalanceado, uma vez que o tratamento conjunto dessas duas questões foi pouco explorado [24, 28]. A revisão sistemática de literatura [29, 30, 8, 31, 32] é utilizada para compreender como esses dois temas estão inter-relacionados, ajudando a identificar técnicas para lidar com desbalanceamento, desvio de conceito, lacunas e como esses temas podem evoluir juntos. Devido a necessidade de uma base de dados adequada de indicadores econômico-financeiros com marca temporal para avaliação de modelos preditivos em fluxo de dados, propõe-se uma base de dados de 84 indicadores a partir de dados de empresas listadas fornecidos à CVM.

1.1 Justificativa

Devido ao impacto nos *stakeholders* das empresas, a predição de DF tem ganhado grande importância econômica. Assim, a antecipação de cenários de DF tem sido estudada em economia, contabilidade e ciências da decisão, sendo motivo de discussão entre a literatura acadêmica e pesquisadores profissionais em todo o mundo. Tanto que, entre 2008 e 2018, foram publicados em média aproximadamente 12 artigos/ano sobre o assunto [32]. Diferentes abordagens tradicionais foram sugeridas com base em testes de hipóteses, mo-

delagem estatística e modelos de AM. Em 2017, a superioridade dos modelos de AM é reconhecida [7].

Apesar dos avanços conquistados nesses anos, ainda é uma área com desafios a serem superados. Normalmente, a DF não é afetada por um único fator, mas é causada pela combinação de vários indicadores financeiros. Além disso, o modelo preditivo precisa estar preparado para operar em um ambiente instável (não estacionário), onde mudanças podem impactar no ambiente de negócio da empresa e consequentemente no resultado das predições, ocasionando a degradação de desempenho. Portanto, alguns desafios são:

- Poucas bases de dados abertas de indicadores econômico-financeiros de empresas com marca temporal;
- Construção de uma base de dados de indicadores econômico-financeiros para predição de DF;
- Desequilíbrio entre a quantidade de empresas operando em normalidade e aquelas em DF;
- Dinamicidade dos ambientes reais com constante geração de informação e repletos de incertezas, tais como: bolhas econômicas [33], eventos naturais, acidentes [34], inflação, variação cambial, pandemia [35].

Por ser um dos fatores responsáveis pela degradação dos modelos preditivos, o desbalanceamento de classes deve ser tratado. Entretanto, a maioria das pesquisas assume fluxos de dados desbalanceados como sendo relativamente equilibrados [24], o que não é a realidade para o problema exposto. Deixando assim, em aberto, a predição de DF em fluxo de dados desbalanceado, que este estudo pretende explorar.

1.2 Objetivo

O objetivo principal da pesquisa é, a partir de dados desbalanceados fornecidos pelas empresas à CVM, minimizar o impacto do desbalanceamento da base de dados e melhorar o desempenho de modelos preditivos no decorrer do tempo. Os dados serão utilizados para elaborar modelos preditivos capazes de identificar empresas em DFs com antecedência de 2, 4, 8, 12, 16, 20 e 24 trimestres. Para isso será necessário:

- Efetuar levantamento bibliográfico (revisão sistemática de literatura);
- Elaborar uma base de dados de indicadores econômico-financeiros com marca temporal;
- Identificar técnicas para predição de DF;

- Identificar técnicas para tratamento de desbalanceamento de classes;
- Identificar técnicas para tratamento de desvio de conceito;
- Realizar experimentos.

1.3 Hipótese

A hipótese deste trabalho é: Em um contexto de fluxo de dados, algoritmos de balanceamento de classes podem ser utilizados para melhorar o desempenho de modelos preditivos de Aprendizado de Máquina para identificar situações de Dificuldade Financeira em base de dados desbalanceada da CVM. Também tenta-se validar as seguintes hipóteses secundárias:

- Dados contábeis fornecidos pelas empresas podem ser utilizados para predição de Dificuldade Financeira;
- Informações fornecidas trimestralmente podem ser utilizadas no aprendizado em fluxo de dados;
- O desbalanceamento de dados trimestrais impacta negativamente na identificação de Dificuldade Financeira;
- O reequilíbrio das classes melhora o resultado de predições do modelo.

1.4 Organização do documento

No Capítulo 2 é apresentado o método de revisão sistemática com o resultado da pesquisa bibliográfica e levantamento do estado da arte de problemas que precisam ser tratados. A primeira seção aborda como a revisão sistemática foi planejada, seguida pela seção de desenvolvimento que descreve como ela foi aplicada. Por fim, são apresentados os resultados obtidos da revisão. No Capítulo 3, de fundamentos, são apresentadas definições importantes para o entendimento do problema e de conceitos que serão tratados posteriormente, entre eles: fluxo de dados e seus desafios, aprofundando, na segunda seção, sobre desbalanceamento; e termina explicando a questão da DF e a importância de predizer essa situação. No Capítulo 4 é detalhada uma solução para resolver o problema do uso de modelos estáticos na predição de dificuldades financeiras de empresas em que os dados apresentam-se como um fluxo de dados, com problemas de desbalanceamento e evolução no decorrer do tempo. No Capítulo 5, são apresentados os resultados alcançados através de experimentos combinando diferentes possibilidades de configuração. Ao

final, no Capítulo 6, são apresentadas as conclusões deste estudo, identificando o que não está no escopo deste trabalho, as limitações e as lacunas que podem ser desenvolvidas em trabalhos futuros. Além disso, apresenta alguns os produtos desenvolvidos durante o período desse estudo, como base de dados de indicadores e artigos que foram elaborados para publicação em conferências e em revistas.

Capítulo 2

Revisão sistemática

Neste estudo foi adotado o método de revisão sistemática de literatura, utilizado em levantamentos bibliográficos, pois é menos suscetível a viés [29]. Esse método permite catalogar e avaliar estudos relevantes para produção de conteúdo acadêmico [36]. Por fim, por ser confiável, rigorosa e auditável é possível reproduzi-la em novos estudos, permitindo a validação por outros pesquisadores.

Quando aplicada, a revisão sistemática costuma ser dividida em três etapas: planejamento, desenvolvimento e resultados [37]. No planejamento é definido o protocolo de revisão, as questões de pesquisa, os critérios de inclusão, os critérios de exclusão e um sistema de pontuação para classificar os artigos. O desenvolvimento é a etapa em que o planejamento é executado e, caso necessário, o planejamento pode ser revisitado. Por fim, no resultado, são apresentadas as lacunas de pesquisas anteriores, os jornais/revistas mais relevantes na área estudada e os conhecimentos que podem ser aproveitados e/ou aprimorados.

A busca por referências ocorreu em dois momentos. Primeiramente, para entender como identificar se uma empresa está em uma situação de DF e, assim, prever uma possível falência. Nesse momento foi observada a utilização de indicadores econômico-financeiros para essa finalidade, que são atualizados periodicamente em intervalos anuais ou trimestrais e a existência de um forte desbalanceamento entre classes. A existência de atualização periódica dos indicadores caracteriza os dados como não-estacionários. Essa característica levou a pesquisa ao segundo momento, quando foi necessário pesquisar sobre fluxos de dados e como tratar do desbalanceamento de classes durante o processo de predição de DF.

O restante deste capítulo está dividido nas etapas da revisão sistemática. Na Seção 2.1, de planejamento, é apresentado o protocolo de revisão, as questões de pesquisa e a estratégia de busca do estudo. A Seção 2.2, de desenvolvimento, detalha como foi aplicado o protocolo para seleção dos artigos e a quais informações relevantes foram extraídas dos

mesmos. Ao final, na Seção 2.3, de resultados, são apresentadas as conclusões obtidas a partir da revisão.

2.1 Planejamento

Em pesquisa recente foram encontradas quatro revisões de literatura relevantes publicadas nos anos de 2019 e 2020. A revisão publicada por Alam, T. *et al.* (2020) [38], abordou a questão da falência de empresas a partir de 2016. Eles utilizaram o estudo feito por Barboza, F. *et al.* (2017) [7] comparando técnicas de AM *i.e.* Máquina de Vetores de Suporte (*Support Vector Machine* - SVM), *bootstrap aggregating* (*bagging*), *boosting* (AdaBoost), *Random Forest* (RF), Redes Neurais Artificiais (RNA) e RL com a análise de discriminante (método estatístico) proposto por Altman, E. (1968) [3], quando foi verificada a superioridade das técnicas de AM. A partir desse ponto, eles evidenciam uma forte tendência de pesquisas em AM e listam ao menos nove países objeto de estudos nessa área. Por fim, destacam a importância de uma preparação adequada dos dados e o tratamento de questões como: normalização, valores faltantes, transformação dos dados e desbalanceamento.

Barboza *et al.* (2020) [32] aborda a predição de DF. Nessa revisão foram analisadas 165 publicações, entre 1992 e 2019, por meio do método *Knowledge Development Process - Constructivist* (ProKnow-C). Shi & Li (2019) [39] apresenta um gráfico anual de publicações, de 1991 até 2018, que complementado com dados deste estudo, para os anos de 2019, 2020, 2021 e 2022, produzem o gráfico apresentado na Figura 2.1, onde o eixo x representa os anos, o eixo y a quantidade de artigos e a linha de tendência vermelha ilusta o crescente interesse pela área.

Na Figura 2.1, é possível observar a evolução anual e o crescimento acentuado no número de publicações a partir do ano de 2009, com uma média de aproximadamente 32 publicações por ano, alcançando um pico de 38 publicações em 2011, 2017 e 2022. Anualmente, a partir de 2008, a quantidade estudos publicados utilizaram técnicas de IA passa a ser maior que a quantidade de estudos utilizando outras técnicas. Os jornais que concentraram o maior número de publicações foram *Expert Systems With Applications*, *Knowledge-Based Systems* e *European Journal of Operational Research* com mais de dez publicações em cada, com destaque para o primeiro, que concentrou 57 publicações.

Os estudos tratam de dados estacionários, porém, as informações financeiras de empresas seguem uma periodicidade de divulgação. Cada divulgação recebe uma marca de tempo (*timestamp*) seguindo uma ordem temporal contínua e tendendo ao infinito, pois essas informações devem ser fornecidas em todo tempo de vida de cada empresa. Portanto, pelo fato de informações financeiras serem dados não-estacionários, é necessário

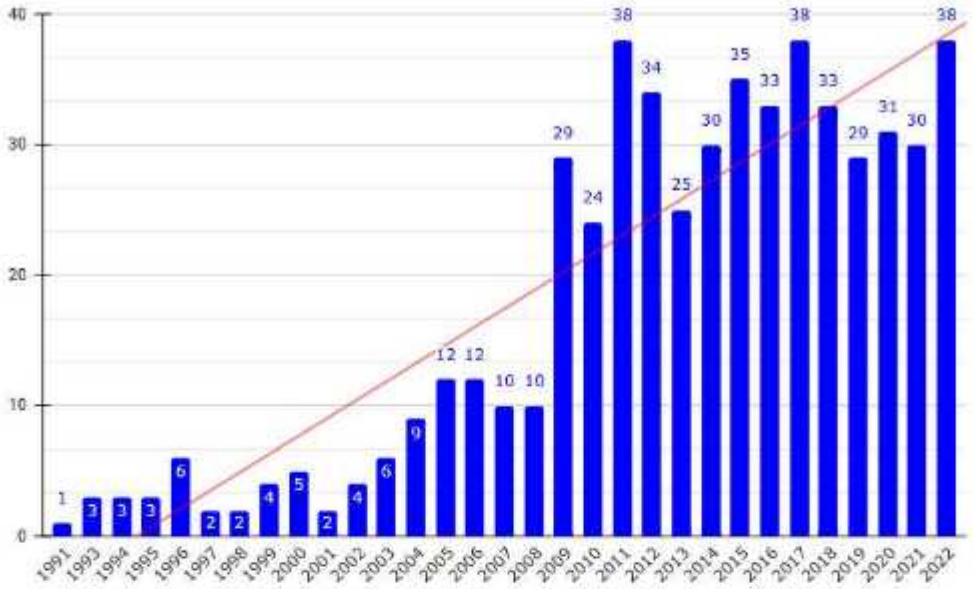


Figura 2.1: Evolução anual de publicações sobre predição de DF/falência.

compreender o funcionamento de fluxo de dados e como modelos preditivos podem ser treinados nesse contexto.

Em Gome, H. M. *et al.* (2019) [1], os autores trazem informações sobre o estado da arte, desafios e oportunidades relacionadas a fluxos de dados. Comumente conhecido por aprendizado incremental, aprendizado *online* ou aprendizado em fluxo de dados. Eles destacam a importância de quatro etapas consecutivas: o pré-processamento dos dados, o processo de aprendizado, a auto-avaliação e adaptação. A primeira trata os atributos para resolver problemas de valores inválidos (valores nulos, ruídos, formatos desconhecidos, etc), redução de dimensionalidade e seleção de atributos. A segunda preocupa-se com a rotulagem parcial ou atrasada, o aprendizado por comitê, o desbalanceamento de dados e a detecção de anomalias. A terceira procura identificar o momento de início da degradação do modelo e suas causas. Por fim, a quarta adapta o modelo à mudança de conceito.

Wang, S. *et al.* (2018) [28] consideram dois problemas inerentes ao fluxo de dados: o desbalanceamento e o desvio de conceito. Ambos presentes, normalmente de forma conjunta, em problemas de AM em fluxos. Os autores destacam que apesar dessa combinação de problemas existir de maneira frequente em situações reais, poucos estudos tratam essa questão e propõem: (*i*) um *framework* para tratamento desses casos; (*ii*) alguns algoritmos para minimizar esses problemas de forma conjunta. Além disso, destaca a falta de estudos para avaliar os efeitos do desbalanceamento de dados em desvio de conceitos.

Buscando dar continuidade a esses estudos foi realizado um levantamento bibliográfico complementar, compreendendo o período de 2019 a 2022. Esse levantamento se justifica, pois trata-se de um período considerável em termos de avanços e inovações, especialmente

Tabela 2.1: Expressões utilizadas para buscar por publicações.

Tema	Ref.	Expressão de busca
Predição de DF	1	predict* AND (insolve* OR bankrupt*) AND financial AND "machine learning"
Fluxo de dados	2	("concept drift" OR "concept feature") AND "data stream" AND "machine learning"
	3	"??balance?" AND "stream*" AND "machine learning"

em uma área que tem atraído tanto interesse da academia e da indústria[38]. Assim, foi elaborada a seguinte pergunta de pesquisa:

Como tratar o desbalanceamento na base de dados da CVM em um cenário de fluxo de dados?

A questão de pesquisa deu origem às expressões de busca, que foram organizadas em temas. Tema 1, aplicou-se a expressão (1) da Tabela 2.1, para pesquisar sobre a questão de DF e falência. Tema 2, expressões (2) e (3) da Tabela 2.1, para pesquisar sobre desbalanceamento e desvio de conceito em fluxo de dados. Todas foram utilizadas no portal de periódicos da CAPES[1].

Por meio do portal de periódicos as expressões foram aplicadas em 295 bases de conteúdo científico, abrangendo 66 da áreas de ciências da computação. Entre as bases consultadas estão:

- ArXiv.org - Cornell University (`https://arxiv.org/`)
- SCOPUS (`https://www.scopus.com`)
- SciELO - Web of Science (`https://www.webofknowledge.com`)

Diante do grande número de artigos retornados, foi necessário definir critérios de inclusão e de exclusão para seleção das publicações mais relevantes. Os critérios foram divididos em dois grupos: critérios globais de inclusão e critérios específicos de exclusão.

Os critérios globais de inclusão de publicações (aplicáveis aos dois temas) foram:

1. Escritas em inglês;
2. Revisão por pares;
3. Posteriores a 2018.

Os critérios de exclusão foram definidos para cada momento da pesquisa, dividido em dois temas. O primeiro, para conhecer sobre o estado da arte de predição de DF e, o segundo, para identificar a melhor forma de lidar com o desbalanceamento em fluxo de dados.

[1]https://www-periodicos-capes-gov-br.ezl.periodicos.capes.gov.br/index.php

1. **Tema 1**: Predição de falência e/ou DF (expressões de busca: 1)
   (a) Não usa técnicas de AM;
   (b) Não trata de predição de falência de forma ampla, mas de setores específicos (agrícola, governo, construção civil, mineração, etc.);
   (c) Não trata de predição de DF ou falência;
   (d) Não trata de predição de falência para empresas;
   (e) Não utiliza indicadores financeiros para prever falência (*i.e.* utiliza análise de documentos ou notícias ou redes financeiras).
2. **Tema 2**: AM em fluxo de dados, com desbalanceamento (expressões de busca: 2 e 3).
   (a) Não faz uso de técnicas de AM no contexto de fluxo de dados;
   (b) Não faz uso de técnicas de desbalanceamento;
   (c) Utiliza técnicas de aprendizado em áreas específicas (saúde, agricultura,monitoramento de água, etc);
   (d) Não lida com problemas de classificação;
   (e) Trata de desbalanceamento, mas não no contexto de fluxo de dados.

Após consultar as publicações, através do portal de periódicos da CAPES, os critérios de exclusão foram aplicados para excluir aqueles artigos de baixa relevância. Então, os artigos resultantes foram classificados para permitir ordenar os artigos de melhor qualidade para os de menor qualidade. A pontuação é decorrente de uma lista de itens desejáveis. Caso o estudo atendesse o item de maneira integral ele pontuava com 1, caso atendesse de maneira parcial ele pontuava com 0,5 e caso não atendesse o item ele não pontuava. O somatório dos pontos de cada item resultou em um valor de qualidade do artigo. Os itens considerados foram:

1. O objetivo do trabalho está alinhado com o objetivo desta pesquisa?
2. Os desafios de resolver o problema de desbalanceamento foram apresentados?
3. A base de dados está disponível para uso?
4. Os atributos dos dados foram explicados?
5. Os detalhes da abordagem permitem replicá-la?
6. Os resultados foram comparados a outros estudos na área?
7. As limitações do estudo foram descritas?

8. O código fonte está disponível?
9. Os hiperparâmetros dos algoritmos foram descritos?

Antes da leitura das publicações selecionadas foram definidas quais informações precisavam ser obtidas, a saber:

- Informações sobre as publicações (título, autor, ano da publicação, jornal e fator de impacto)
- Fatores de desempenho que foram utilizados.
- Os algoritmos utilizados.
- Quantidade de empresas utilizada no estudo.
- Atributos utilizados para treinar o modelo.
- Base de dados utilizada no estudo.
- Observações.

2.2 Desenvolvimento

Na Figura 2.2 é possível observar a sequência de passos executados durante a revisão sistemática. Primeiramente, foi executada a Atividade 1.1, em Tema 1, utilizando a expressão de busca (1) definida na Tabela 2.2. Na Atividade 1.2, o resultado da pesquisa foi filtrado segundo os critérios de exclusão do Tema 1, resultando em 121 publicações. Em um outro momento, foi executada a Atividade 2.1, em Tema 2, utilizando as expressões de busca (2 e 3) definida na Tabela 2.2. Após unir os resultados das Atividades 1.2 e 2.1, obteve-se um total de 878 artigos. Na Atividade 2.2, esse resultado foi filtrado segundo os critérios de exclusão do Tema 2, obteve-se 12 artigos, após exclusão de duplicações, restaram 11 artigos sobre desbalanceamento, desvio de conceito e FDP em fluxo de dados.

A aplicação dos critérios de exclusão e a quantidade de artigos excluído por cada item da lista de critérios de exclusão pode ser verificada em Tabela 2.2. Os critérios do Tema 1 foram abreviados como 1.a, 1.b, 1.c, 1.d e 1.e; e, os critérios do Tema 2 foram abreviados como 2.a, 2.b, 2.c, 2.d e 2.e.

2.3 Resultado

Nos últimos 15 anos, a IA se consolidou como a principal ferramenta para predição de falência. Essa posição deve-se aos resultados obtidos por meio de estudos que atingiram

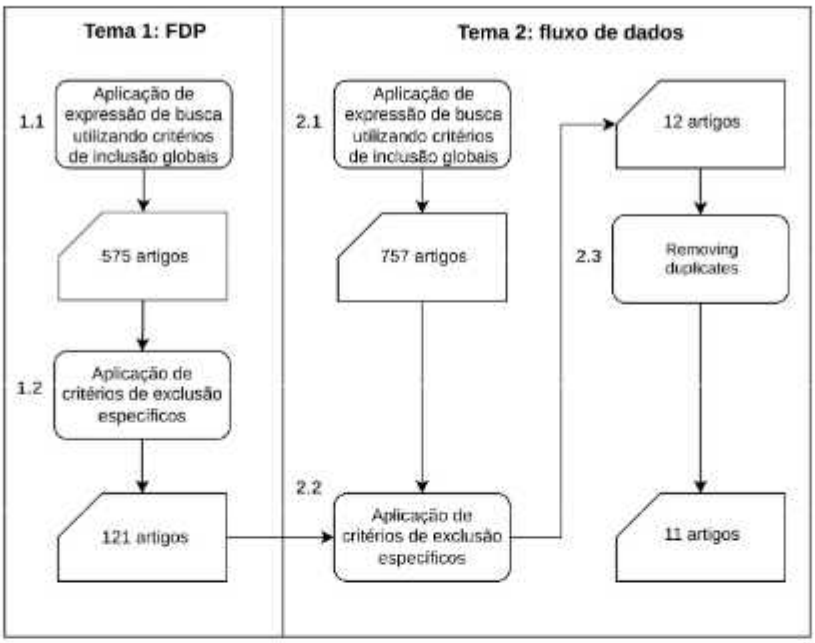


Figura 2.2: Fluxo de seleção de publicações.

Tabela 2.2: Quantidade de publicações obtidas e filtradas.

Atividade	**Resultado**
1.1 Pesquisa de artigos (tema 1)	575
1.2 Aplicação dos critérios de exclusão	-368 (1.a) -10 (1.b) -58 (1.c) -4 (1.d) -14 (1.e)
Resultado A (1.1 + 1.2)	*121*
2.1 Pesquisa de artigos (tema 2)	757
Resultado B (A + 2.1)	*878*
2.2 Aplicação dos critérios de exclusão	-578 (2.a) -154 (2.b) -39 (2.c) -45 (2.d) -50 (2.e)
Resultado C (B + 2.2)	*12*
2.3 Remoção de duplicações	-1
Resultado Final (C + 2.3)	**11**

acurácia superior a 90% [38, 31, 40]. Entretanto, esses estudos utilizaram dados estacionários, tratando o ambiente como estático. Isso pode ser afirmado porque as fontes de dados trazem informações das empresas em um momento específico. Elas não têm marca temporal e, portanto, não consideram as informações de momentos anteriores. Alam, T. *et al.* (2020) [38] utilizou dados de empresas da Polônia enquanto que Wang, H. & Liu, X. (2021) [41] optaram por dados de empresas de Taiwan[2]. Porém, ambos utilizaram dados estáticos de empresas de capital aberto, que devem fornecer as informações a órgãos reguladores de mercado. Algo semelhante ocorre no Brasil, onde a CVM recebe os dados dessas empresas periodicamente. A partir da evidência de que os dados são desbalanceados e de natureza temporal, verificou-se que apenas três estudos tratava a questão da predição de falência e desbalanceamento das classes em um ambiente de fluxo de dados, como feito por Sun *et al.* (2017) [42], Sun *et al.* (2019) [43] e Shen, F. *et al.* (2020) [44].

As revistas/jornais foram catalogadas para identificar a relevância delas no meio acadêmico. Com isso foi possível identificar as áreas de pesquisa de cada jornal em que foi publicado artigos sobre o assunto.

- A1: Ciências da computação (IA)
- A2: Ciências da computação (Aplicações em Ciências da Computação)
- A3: Ciências da computação (Sistemas de Informação)
- A4: Ciências da computação (Diversos)
- A5: Economia, Econometria e Finanças
- A6: Multidisciplinar

As informações foram compiladas na Tabela 2.3, com nome da revista/jornal, quartil de relevância da revista (Qtl), quantidade de artigos publicados sobre o assunto (Pub), área de pesquisa (Área), cite score (Score), rank e SCImago Journal Rank (SJR). Os dados da tabela vieram do Scimago Journal & Country Rank[3] e foram ordenados por quartil e publicações. Para construir esta tabela, foram utilizados os artigos retornados após a Atividade 1.1 na Figura 2.2, totalizando 121 artigos. Eles foram filtrados por quartil igual a Q1, Q2 ou Q3 com pelo menos 2 artigos publicados.

[2]https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction
[3]`https://www.scimagojr.com/journalrank.php`

Tabela 2.3: Revistas com duas ou mais publicações, classificadas nos quartis Q1, Q2 e Q3.

Revista/Jornal	**Qtl**	**Pub**	**Área**	**Score**	**Rank**	**SJR**
Sustainability (Basel, Switzerland)	Q1	10	A_4	5	15/90	0,664
Expert systems with applications	Q1	9	A_1	12.2	23/269	2,070
IEEE access	Q1	5	A_4	6,7	34/231	0,927
Knowledge-based systems	Q1	4	A_1	12	25/269	2,192
European journal of operational research	Q1	4	A_4	10.5	16/231	2,354
Computational intelligence and neuroscience	Q1	3	A_4	3,9	63/231	0,863
Computational economics	Q1	3	A_5	3,3	40/189	0,454
Information sciences	Q1	2	A_2	12,1	35/747	2,290
Sensors (Basel, Switzerland)	Q1	2	A_3	6,4	77/353	0,803
Journal of international financial markets, institutions and money	Q1	2	A_5	6,2	27/299	1,310
Journal of international studies	Q1	2	A_5	4,3	139/696	0,427
PloS One	Q1	2	A_6	5,6	15/120	0,852
Journal of forecasting	Q2	4	A_2	3,7	304/747	0,594
Applied sciences	Q2	4	A_2	3,7	301/747	0,507
Complexity (New York, N.Y.)	Q2	3	A_4	2,6	41/90	0,394
Risks (Basel)	Q2	3	A_5	2,2	62/189	0,398
Cogent economics and finance	Q2	2	A_5	2,3	121/299	0,411
Journal of business economics and management	Q2	2	A_5	3,3	206/696	0,485
International journal of finance and economics	Q2	2	A_5	2,1	134/299	0,424
Journal of risk and financial management	Q3	13	A_5	0,6	-	-

Através dessa revisão, a falta de trabalhos com empresas brasileiras ficou evidente. Nos trabalhos relativos ao Tema 1 (falência e DF), observou-se o uso de bases de dados de diferentes localidades. Clement, C. (2020) [31] lista alguns países em que as empresas foram objeto de estudos de predição de falência, a saber: Taiwan, Polônia, França, EUA, Bélgica, Itália, Espanha, China, Hungria, Sérvia, Estônia, Grécia, Índia e Japão. A falta de trabalhos voltados para o Brasil estimulou Barboza, F. *et al.* (2021) [40] a produzir um trabalho sobre empresas da América Latina obtidos na Economatica[4], onde foi utilizado o *Extreme Gradient Boosting* (XGBoost) para predição de DF.

A lista de países de Clement, C. (2020) [31] foi complementada com informações dos anos de 2019 a 2022 e pode ser vizualizada na Figura 2.3, onde é possível observar a frequência em que alguns países foram objeto de estudos sobre DF/falência. Alguns deles, mesmo com a economia menor que a do Brasil, apresentaram um maior número de estudos. Por exemplo, Taiwan e Polônia com 8 e 15 publicações, respectivamente. Esses números podem ser explicados pelo fato das bases de dados estarem disponíveis no repositório da UCI para acesso público.

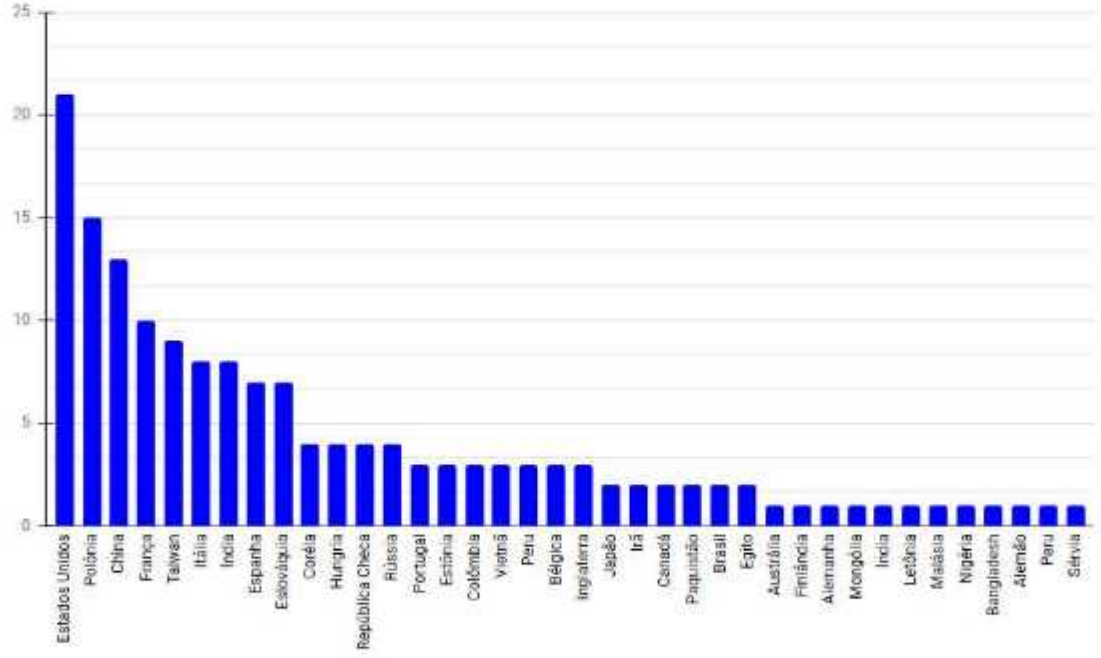


Figura 2.3: Distribuição de publicações sobre DF/falência por localidade.

A maioria dos estudos sobre predição de DF utilizou dados privados, como Compustat[5]: banco de dados contendo dados financeiros e de preços fundamentais de empresas ativas e inativas de capital aberto dos Estados Unidos [7]; Economatica[6]: conjunto

[4]https://economatica.com/

[5]`https://www.library.hbs.edu/find/databases/compustat`

[6]`https://economatica.com/`

de dados com dados do mercado de ações do Brasil, América Latina e Estados Unidos [40]; *AIDA do Bureau van Dijk*[7]: contém informações abrangentes sobre empresas na Itália [45]; *China Stock Market & Accounting Research* (CSMAR)[8]: é um banco de dados abrangente orientado para pesquisa com foco em Finanças e Economia da China [44, 46, 47], e; Orbis[9]: banco de dados que contém informações sobre cerca de 450 milhões de empresas e entidades em todo o mundo [48, 49]. Esses dados geralmente contêm informações extensas que podem ser organizadas cronologicamente, isto é por ano, semestre ou trimestre, representando assim um fluxo de dados. No entanto, os dados não estão prontos para uso e precisam de mais computação para extrair os atributos (indicadores econômico-financeiros). Além disso, a maioria dos estudos considerou esses dados como estacionários.

Por outro lado, alguns conjuntos de dados estão disponíveis gratuitamente em repositórios públicos ou pessoais, eles contêm indicadores econômico-financeiros e estão prontos para uso. No entanto, não levam em consideração a questão temporal e, portanto, não podem ser tratados como fluxos de dados. Alguns deles estão disponíveis no repositório de AM da UCI[10], é o caso dos dados de empresas polonesas [50] e dados de empresas taiwanesas [51], do repositório OpemML[11], do Kaggle[12] e dos dados de previsão de falência de empresas americanas no mercado de ações em um repositório pessoal[13] [52] .

Todos os trabalhos fizeram uso de indicadores econômico-financeiros. Todavia, alguns tentaram agregar mais informações a análise e trabalharam com indicadores de governança corporativa [53], o compartilhamento de gestores e diretores entre empresas [54], dados textuais apresentados em documentos de divulgação de informação relevante aos investidores [55] ou informações de impostos atrasados [56]. Apesar dos indicadores econômico-financeiros estarem presentes em todos os estudos, não existe um conjunto bem definido dos indicadores de maior relevância, pois há estudos com 6 indicadores até estudos com 263 indicadores [31].

Quanto ao desbalanceamento em fluxo de dados, foram utilizadas diferentes técnicas que podem ser classificadas em ao menos três opções: abordagem de pré-processamento (sobreamostragem e/ou subamostragem), abordagem em algoritmo e abordagem de aprendizado em comitê. A Tabela 2.4 apresenta algumas dessas técnicas ordenadas de forma crescente pela frequência da abordagem. Ao todo, foram usadas nove técnicas da abordagem de pré-processamento, duas técnicas da abordagem em algoritmo e quatro técnicas

[7] `https://aida.bvdinfo.com`
[8] `http://cndata1.csmar.com/`
[9] `https://www.bvdinfo.com/en-gb/our-products/data/international/orbis`
[10] `https://archive.ics.uci.edu/ml/index.php`
[11] `https://www.openml.org/`
[12] `https://www.kaggle.com/`
[13] `https://github.com/sowide/bankruptcy_dataset`

Tabela 2.4: Algoritmos para tratar desbalanceamento em fluxo de dados.

Abordagem	**Nome**	**Ref.**
Em algoritmo	CALMID	[57]
	CSDS	[58]
Aprendizado em comitê	AWE and AUE	[59]
	DUE	[24]
	Adapted Neural Network	[60]
	WEOB1 e WEOB2	[61]
Pré-processamento	EasyEnsemble	[62]
	SMOTE	[63]
	REA	[64]
	ERDDM	[65]
	Unrolled GAN	[66]
	Novel version of SMOTE	[13]
	AWSMOTE	[67]
	MIS-ELM	[68]
	ANS-REA	[44]

da abordagem em comitê. Entre as técnicas de pré-processamento há uma predominância na técnica de SMOTE e suas variantes, totalizando 4 exemplos: SMOTE, *Novel version of SMOTE*, *Adaptive-Weighting SMOTE* (AWSMOTE) e *Adaptive Neighbor SMOTE-Recursive Ensemble Approach* (ANS-REA).

Um dos trabalhos disponibilizou o experimento no *Microsoft Azure Machine Learning Studio* [38]. Em pesquisas na Web, foi possível encontrar, no Kaggle[14], código fonte de um projeto de predição de falência de empresas que utiliza dados de empresas de Taiwan[15], com tratamento de desbalanceamento da base de dados através de SMOTE e utilizando quatro algoritmos distintos de AM, a saber: RL, *Random Forest*, XGBoost [40, 69] e CatBoost (*Categorial Boosting*), que obtiveram acurácia de 84%, 97%, 95% e 97%, respectivamente, com os dados sendo processados de maneira estacionária.

Outras técnicas também têm demonstrado resultados promissores. Alaka H. *et al.* (2018) [8] citam *Support Vector Machine* (SVM), RNA, Árvore de Decisão (AD), Algoritmo genético (AG), Raciocínio Baseado em Casos (RBC) e Korol, T. (2019) [4] apresenta *fuzzy set* como uma alternativa. Normalmente, a escolha da técnica vai depender do contexto do usuário, podendo variar caso ele seja um investidor, cliente, proprietário, agência de governo ou auditor [8]. Esses modelos foram utilizados para tentar classificar as empresas em duas classes DF e Sem Dificuldade Financeira (SDF), antecipando o evento em um período de 1 a 3 anos [70], 1 a 5 anos [69] e, em alguns casos [4], de 1 a 10 anos. Nesses estudos os dados foram tratados como estacionários. Entretanto, o ambiente é dinâmico

[14]https://www.kaggle.com/code/marto24/bankruptcy-detection
[15]https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

podendo ocorrer choques econômicos que interferem nas empresas e geram impacto nos indicadores.

Avaliar o desempenho desses modelos é importante para identificar em que situação e com que conjunto de dados cada um deles pode ser utilizado a fim de obter um melhor desempenho. Para isso, no geral, os estudos que lidam apenas com a predição de falência fazem uso da acurácia. Porém, quando consideramos a questão do desbalanceamento dos dados a acurácia deixa de ser uma boa métrica [24]. Assim, é necessário o uso de métricas adequadas, como por exemplo: área abaixo da curva ROC [24, 44, 57, 64, 69], sensibilidade (revocação, *Recall*, Taxa de Verdadeiro Positivo (TVP)) [24, 57, 60, 64], média geométrica (*geometric mean*, *G-mean*) [24, 57, 60], medida F ($F_measure$, F_1) [24, 44], Kappa [44], precisão [24] e especificidade (*specificity*, Taxa de Verdadeiro Negativo (TVN)). A sensibilidade pode ser utilizada para avaliar o desempenho do modelo no conjunto de dados minoritários e a precisão é a proporção entre os exemplos de classe minoritária que são classificados corretamente e os exemplos previstos como classe minoritária [24].

Por fim, existem alguns *frameworks* ou bibliotecas de AM que podem ser aplicadas na pesquisa acadêmica e no desenvolvimento de aplicações comerciais, a saber: River[16], MOA[17] (*Massive Online Analysis*), ADAMS[18] (*Advanced Data mining And Machine learning System*), SAMOA[19] (*Scalable Advanced Massive Online Analysis*), VW[20] (*Vowpal Wabbit*), StreamDM[21], Scikit-multiflow[22] [71] e Ray RLlib[23].

[16]https://riverml.xyz/
[17]http://moa.cms.waikato.ac.nz
[18]https://adams.cms.waikato.ac.nz
[19]http://samoa.incubator.apache.org
[20]https://github.com/VowpalWabbit/vowpal_wabbit
[21]http://huawei-noah.github.io/streamDM/
[22]https://scikit-multiflow.github.io/
[23]https://ray.readthedocs.io/en/latest/rllib.html

Capítulo 3

Fundamentação teórica

Há muito tempo o ser humano vem se organizando como sociedade e buscando formas de facilitar o comércio. Primeiramente, passando da economia baseada em trocas diretas (escambo) à economia baseada em trocas indiretas, ou economia monetarizada, onde se utiliza um objeto intermediário para efetuar as trocas, a moeda [72]. Isto é, um objeto com valor monetário, como: conhas, sal, ouro ou prata.

Quando as trocas são feitas por meio de moeda, acabam criando outras necessidades, por exemplo, onde guardar esses valores. Assim, em 1406, é criado primeira instituição de crédito moderna (banco) *Casa di San Giorgio*[73]. Da criação do primeiro banco até os dias de hoje esse mercado tem passado por constantes evoluções, por momentos de grandes crescimentos econômicos e por grandes recessões. Atualmente, os bancos desempenham um importante papel em todas as sociedades humanas, onde são fundamentais para o crescimento de economias locais ou responsáveis por fortes recessões [74].

Com o advento dos meios digitais, por meio dos bancos, as economias têm se conectado para realizar o comércio global e em países emergentes, como o Brasil, o investimento extrangeiro representa um importante motor da economia, gerando empregos e melhorando a infraestrutura local. São os bancos e os grandes fundos de investimentos que fazem chegar os recursos dos investidores às economias locais [75]. Entretanto, toda essa conexão entre as instituições financeiras deve ser monitorada e acompanhada, pois além de benefícios elas podem ser responsáveis por sérias crises econômicas algumas vezes iniciadas em setores específicos da economia [25]. Quando uma crise atinge o setor financeiro, devido a interdependência existente entre as instituições financeiras, alastra-se de forma rápida, grave e acentuada por toda economia local podendo afetar a economia global.

3.1 Risco sistêmico

O setor financeiro está fortemente interconectado, com uma grande quantidade de capital circulando entre suas instituições. Estima-se que 23% dos ativos e 48% dos passivos em instituições bancárias são provenientes de outras instituições financeiras [76]. Essa interdependência permite melhor compartilhamento de riscos e alocação de capital, porém abre caminho para o risco sistêmico, como ficou evidente durante a crise financeira de *subprime* de impacto global, em 2008 [77]. Uma certeza surge dessa crise, esse não foi o único, nem será o último choque econômico pelo qual a economia global terá que passar, basta vermos a história recente. Nos últimos cem anos foram: em 1929, a grande depressão [78]; em 1970, crise dos países da América Latina [79]; em 1985, a bolha imobiliária e das ações no Japão [80]; em 1994, a crise dos mercados emergentes iniciada no México, pelo qual ficou conhecida como *Tequila crisis* [81]; em 2008, o *subprime*; e, mais recentemente, em 2019, a crise desencadeada pela pandemia de Covid-19 [35].

Nesse contexto, é importante entender como setores da economia real podem impactar no risco sistêmico [82]. Devemos observar que os responsáveis por desencadear uma reação sistêmica no sistema financeiro são as grandes corporações financeiras, aquelas consideradas grandes demais para falir (*too-big-to-fail*) [83]. Então, o que pode levar essas grandes instituições financeiras à falência? Um dos principais fatores é o risco financeiro assumido por essas instituições em suas operações de empréstimos a terceiros, especialmente às empresas (grandes tomadoras de empréstimos). Assim, uma taxa de inadimplência além do esperado impacta na liquidez da instituição e em suas obrigações com outras instituições, desencadeando uma reação em cadeia (sistêmica) [25].

É sabido que a falência de uma empresa não é um evento abrupto. Tratando-se de um processo com diferentes fases. Dessa forma, é importante observar além do momento de decretação da falência e ampliar o olhar para o momento em que a empresa apresenta alguma DF. A intenção é evitar o momento de colapso do negócio utilizando dados fornecidos pelas empresas a órgãos reguladores (*i.e.* CVM). Desses dados é possível extrair indicadores econômicos e financeiros para predição da DF [53]. A possibilidade de identificar a DF corporativa tem grande valor, pois proprietários e investidores podem tomar decisões para tentar evitar sua falência ou diminuir as perdas. Na maioria dos casos, a DF não é afetada por um único fator, por isso é necessário utilizar vários indicadores econômico-financeiros.

3.2 Predição de dificuldade financeira

A capacidade de identificar/prever DF permitirá uma melhor análise dos riscos das instituições financeiras a fim de evitar ou diminuir o risco sistêmico. Essa possibilidade proporciona mais segurança e estabilidade aos mercados. Além disso, pode ser utilizada pelas instituições financeiras no momento de avaliação de empresas para concessão de empréstimos ou por investidores auxiliando na escolha de onde investir seus recursos.

Existem, ao menos, dois problemas que interferem na tentativa de prever ou identificar o momento de dificuldade de uma empresa. Primeiro, apesar de especialistas de domínio da empresa determinarem os vários indicadores financeiros, ainda é um desafio a maneira pela qual eles devem ser combinados. Em segundo lugar, o conjunto de dados usado para treinar o modelo é desbalanceado, pois existe uma quantidade maior de empresas saudáveis do que em dificuldade [38].

Devido a importância desse tema para economia local e global, o Fundo Monetário Internacional (FMI) e o Banco Mundial vêm trabalhando em como prever situações de falência [38]. Um dos primeiros trabalhos sobre isso remonta a meados do século XX (1942), que tentava explicar e prever o fracasso de negócios por meio de sinais econométricos [84]. A década de 60 é um divisor de águas nas pesquisas dessa área, pois marca a utilização de ferramentas de estatística para tentar predizer o insucesso de negócios. O trabalho mais relevante nessa década foi o de ADM, onde Altman, E. (1968) [3] buscava avaliar a eficácia da análise de indicadores como uma ferramenta analítica, utilizando como estudo de caso a predição de falência de empresas. Foram utilizados cinco indicadores: liquidez, lucratividade, produtividade, alavancagem e giro de ativos. O modelo proposto avaliava a situação monetária da empresa e organizava o resultado em três classes: perigoso, moderado e seguro.

A partir de dados financeiros de arquivos contábeis (*e.g.* Balanço Patrimonial de Ativos (BPA), Balanço Patrimonial de Passivos (BPP), Demonstração de Resultado (DRE) e Demonstração de Fluxo de Caixa (DFC)), é possível extrair vários indicadores econômico-financeiros, como solvência, estrutura de capital, fluxo de caixa, estrutura de propriedade, liquidez, rentabilidade, volume de negócios, indicadores relacionados à atividade, estrutura financeira foram usados, entre outros. Todavia, os indicadores para previsão de falência corporativa ainda são discutíveis [38]. Por isso, é possível encontrar bases de dados para teste com dezenas de indicadores, como por exemplo a base de empresas da Polônia[1] e a base de empresas de Taiwan[2], com 63 e 95 atributos, respectivamente. Este último, tendo sido utilizado em um trabalho disponibilizado no *Kaggle* [85] (com código fonte) comparando os algoritmos RL, *Random Forest* (RF), *Extreme Gradient Boosting*

[1]https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data
[2]https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

(XGBoost) e *Categorial Boosting* (CatBoost). Além disso, o Portal de dados abertos da CVM é uma fonte primária de informação, pois contém dados enviados diretamente pelas empresas que, por força de lei [5], são obrigadas a fornecer periodicamente informações contábeis seguindo padrões internacionais definidos pelo *International Accounting Standards Board* (IASB) [86].

Nas últimas décadas, o uso de técnicas de IA para tentar predizer situações de estresse em negócios passou a ser mais frequente. Várias técnicas têm sido testadas, desde o uso de Rede Neural Convolucional (RNC) para interpretar gráficos de indicadores financeiros [87] até algoritmos mais tradicionais de IA, como a RL [70]. Barboza, F. *et al.* (2017) [7] reconheceram a superioridade das ferramentas de IA [7]. Desde então, algoritmos como SVM, *Logistic Model Tree* (LMT), AD (J48), RF [38], XGBoost [40], entre outros, vêem demonstrando resultados animadores com acurácia superior a 90% para o problema de classificação binária de empresas através de indicadores econômico-financeiros.

3.3 Fluxo de dados

Quando comparamos as fontes de dados atuais com as da década anterior observamos que elas estão crescendo de maneira acelerada e tornando-se cada vez mais onipresentes [1]. É possível observar celulares nas mãos das pessoas à nossa volta, além de vários outros dispositivos conectados, *e.g.* TVs, geladeiras, ar-condicionado, lâmpadas, entre outros. Todos gerando dados de maneira contínua. Com a implantação do 5G a tendência é que mais dispositivos sejam conectados, coletando ainda mais dados. A academia vem percebendo isso e tem desenvolvido diversos algoritmos de AM para fluxos de dados (aprendizado incremental, mineração de dados em tempo real, análise em tempo real ou aprendizado de fluxo), onde sequências de itens, possivelmente infinitas que chegam continuamente com um *timestamp* associado a cada item [2, 19, 88]. Assim, devido a ordem temporal dos itens, é necessário construir e manter modelos preditivos capazes de avaliar esses itens considerando essa característica, algumas vezes, em tempo real.

Podemos encontrar exemplos de aplicações de aprendizado em fluxo no mundo real, por exemplo, telefones celulares, controles de processos industriais, interfaces de usuário inteligentes, detecção de intrusão, detecção de spam, detecção de fraude, monitoramento do SFN, entre outros [26, 19, 89]. A análise de dados em fluxo está se tornando um padrão para extrair conhecimento útil permitindo que pessoas, organizações e órgãos governamentais reajam rapidamente quando surgem novas tendências. Este estudo busca analisar as empresas que enviam informações financeiras e contábeis à CVM[3] de maneira constante, trimestralmente. Pois, esses dados permitem prever a DF possibilitando a

[3]http://dados.cvm.gov.br/

órgãos governamentais, como Banco Central do Brasil (BCB), atuarem para mitigar ou atenuar o risco sistêmico para garantir a solidez do SFN[4] [25].

Esse cenário nos leva a supor que temos que lidar com um volume de dados crescente, potencialmente infinito, que pode chegar continuamente em lotes de itens ou item a item. Em contraste com os sistemas tradicionais (aprendizado em lote) onde há acesso livre aos dados históricos. Os sistemas de processamento tradicionais, baseados em lote, podem armazenar grandes coleções de dados permitindo que seus usuários executem consultas ou transações, assumindo que os dados estão em repouso ou possibilitando acessos simultâneos [26, 88]. Esses sistemas não costumam ser atualizados continuamente, ao invés, reconstroem regularmente novos modelos a partir do zero. A depender da quantidade de dados e do tipo de modelo de aprendizado, essa estratégia pode implicar em custos computacionais consideráveis [26, 19].

Um fluxo de dados no momento t (*timestamp*), composto de n itens rotulados, pode ser representado da seguinte forma:

$$(x_i^t, y_i^t) \in X^t = \{(x_1^t, y_1^t), (x_2^t, y_2^t), ..., (x_n^t, y_n^t)\}, \text{ onde } 1 \leq i \leq n$$

Onde X é o conjunto de todas as tuplas (x_i, y_i), cada item x_i possui um conjunto de atributos $\{a_1, a_2, ..., a_m\}$ que o caracteriza. O conjunto de atributo é definido da seguinte forma:

$$a_j \in \{a_1, a_2, ..., a_m\}, \text{ onde } 1 \leq j \leq m$$

É necessário adicionar o componente de tempo t (*timestamp*) a cada elemento de x_i. Dessa forma, podemos representar o conjunto de todos os atributos do conjunto X como uma matriz multidimensional com o componente de tempo, assim:

$$x_i^t \in \{(a_1^t, ..., a_m^t)_1, ..., (a_1^t, ..., a_m^t)_n\} \text{ ou } \{a_{1,1}^t, ..., a_{n,m}^t\}$$

Onde, no instante t, um atributo j de um determinado elemento i pode ser representado por $a_{i,j}^t$. O conjunto de rótulos Y pode ser representado como:

$$y_k \in Y = \{y_1, y_2, ..., y_n\}, \text{ onde } 1 \leq k \leq n$$

Também é preciso adicionar o componente de tempo t (*timestamp*) a cada elemento de y_i. Dessa forma, podemos representar o conjunto de todos os atributos do conjunto Y como uma matriz unidimensional com o componente de tempo, assim:

[4]https://www.bcb.gov.br/acessoinformacao/institucional

$$y_k^t \in Y^t = \{y_1^t, y_2^t, ..., y_n^t\}, \text{ onde } 1 \leq k \leq n$$

A possibilidade de mudanças da distribuição dos dados no tempo adiciona complexidade ao processo de treinamento do modelo preditivo [90]. Assim, em momentos distintos, podemos representar a mudança na distribuição de um elemento e seu respectivo rótulo (categoria/classe) da maneira descrita por Gama *et al.* (2014) [18]:

$$\exists x : P^t(x_i, y_k) \neq P^{t+1}(x_i, y_k)$$

Onde $P^t(x, y)$ é a probabilidade de um elemento x_i ser classificado na categoria y_k, em um momento t. Para expandir a compreensão a respeito de desvio de conceito lançando mão da teoria de decisão de Bayes [18], onde uma classificação pode ser descrita pelas probabilidades anteriores das classes $P(y)$ e as funções de densidade de probabilidade condicional de classe $P(X|y)$ para todas as classes y_k. A decisão de classificação é feita de acordo com as probabilidades *a posteriori* das classes, que para a classe y_k podem ser representadas como:

$$P(y_k|x_i) = \frac{P(y_k) \times P(x_i|y_k)}{P(x_i)}, \text{ onde } P(x_i) = \sum_{c=1}^{k} P(y_k) \times P(x_i|y_k)$$

A partir desse ponto, é importante destacar alguns elementos que podem ser impactados [19]:

- $P(y_k)$: probabilidade anterior da classe, mudanças nessa medida afetam a eficiência do modelo e conduzem a um *problema de desbalanceamento*. Evidenciado por uma diferença significativa na quantidade de elementos classificados em cada classe.
- $P(x_i^{t+1}|y_k)$: *probabilidade condicional* de classe, mudanças nessa medida evidenciam o desvio de conceito virtual.
- $P(y_k|x_i^{t+1})$: *probabilidade a posteriori*, alteração neste termo evidencia uma mudança de classificação do elemento x_i^{t+1}, caracterizando o desvio de conceito real.

3.3.1 Aprendizado em fluxo de dados

Uma forma de tratar a chegada contínua de informação é através de uma estratégia de aprendizado incremental ou em blocos (*chunks*). Em inglês conhecido como *stream learning* (aprendizado em fluxo de dados). Essa técnica apresenta vantagens como: mínimo espaço e mínimo tempo de processamento, pois a medida que os dados são recebidos, são tratados e incorporados ao modelo [26, 63, 88, 91, 92, 93, 94]. Devido à sua capacidade de processamento contínuo em grande escala e em tempo real, o aprendizado incremental

ganhou mais atenção no contexto de *Big Data* [11]. Um outra técnica, o aprendizado de fluxo, também apresenta novos desafios e impõe condições rigorosas, como: um pequeno lote de instâncias é fornecida ao algoritmo de aprendizado a cada instante, um tempo de processamento limitado, uma quantidade finita de memória e a necessidade de ter modelos treinados em cada varredura do fluxo de dados [26, 19]. Ademais, os dados podem evoluir ao longo do tempo e ocasionalmente serem afetados por mudanças na distribuição de dados, caracterizando um desvio de conceito [26], forçando o sistema a aprender em condições não estacionárias.

O processo de identificação e tratamento de desvio de conceito pode ser dividido em três fases [19], como ilustrado na Figura 3.1: fase inicial (I), itens obtidos a partir do fluxo de dados são usados para construir o modelo de aprendizado que prevê os valores de destino; fase intermediária (II), o sistema tenta identificar o desvio de conceito nas amostras de dados que estão chegando via fluxo, se não houver desvio na amostra é realizada a previsão das instâncias; e a fase de adaptação (III), ocorre a interpretação e a adaptação do desvio de conceito, em seguida, o mecanismo de esquecimento é realizado. Os retângulos tracejados das fases I (pré-processamento e ajuste do modelo) e III (mecanismo de esquecimento) são passos opcionais.

Na fase intermediária, onde ocorre a detecção do desvio de conceito, várias são as metodologias que podem ser utilizadas para essa finalidade, são conhecidos ao menos oito métodos [19]:

- Método baseado em similaridade e dissimilaridade, utiliza medições de similaridade e dissimilaridade entre a distribuição de amostras de dados em relação ao tempo. Existem vários algoritmos de detecção de desvio de conceito construídos com base nesta abordagem: *Drift Detection Method (DDM)* [95], *Early Drift Detection Method (EDDM)* [96], *Ensemble Classifiers with Drift Detection (ECDD)* [97], *Reactive Drift Detection Method (RDDM)* [98], *Learning with Local Drift Detection (LLDD)* [99], *Dynamic Extreme Learning Machine (DELM)* [100] e *Drift Detection Method with False Positive rate for Multi-label classification (DDM-FP-M)* [101]

- Método estatístico, utilizado para identificar o desvio de conceito comparando o elemento atual com a distribuição histórica dos dados através de testes estatísticos, como média, mediana, curtose, desvio padrão, regressão, teste de hipótese, etc. Medir as semelhanças e diferenças entre as distribuições de itens no decorrer do tempo também é um método adotado. Algumas técnicas conhecidas são: *Statistical Test of Equal Proportions (STEPD)* [102], *Dynamic Clustering Forest (DCF)* [103], *Hoeffding Drift Detection Method (HDDM)* [93], *Fast Hoeffding Drift Detection Method (FHDDM)* [19], *Fisher Test Drift Detector (FTDD)* [104] e *Plover algorithm* [105].

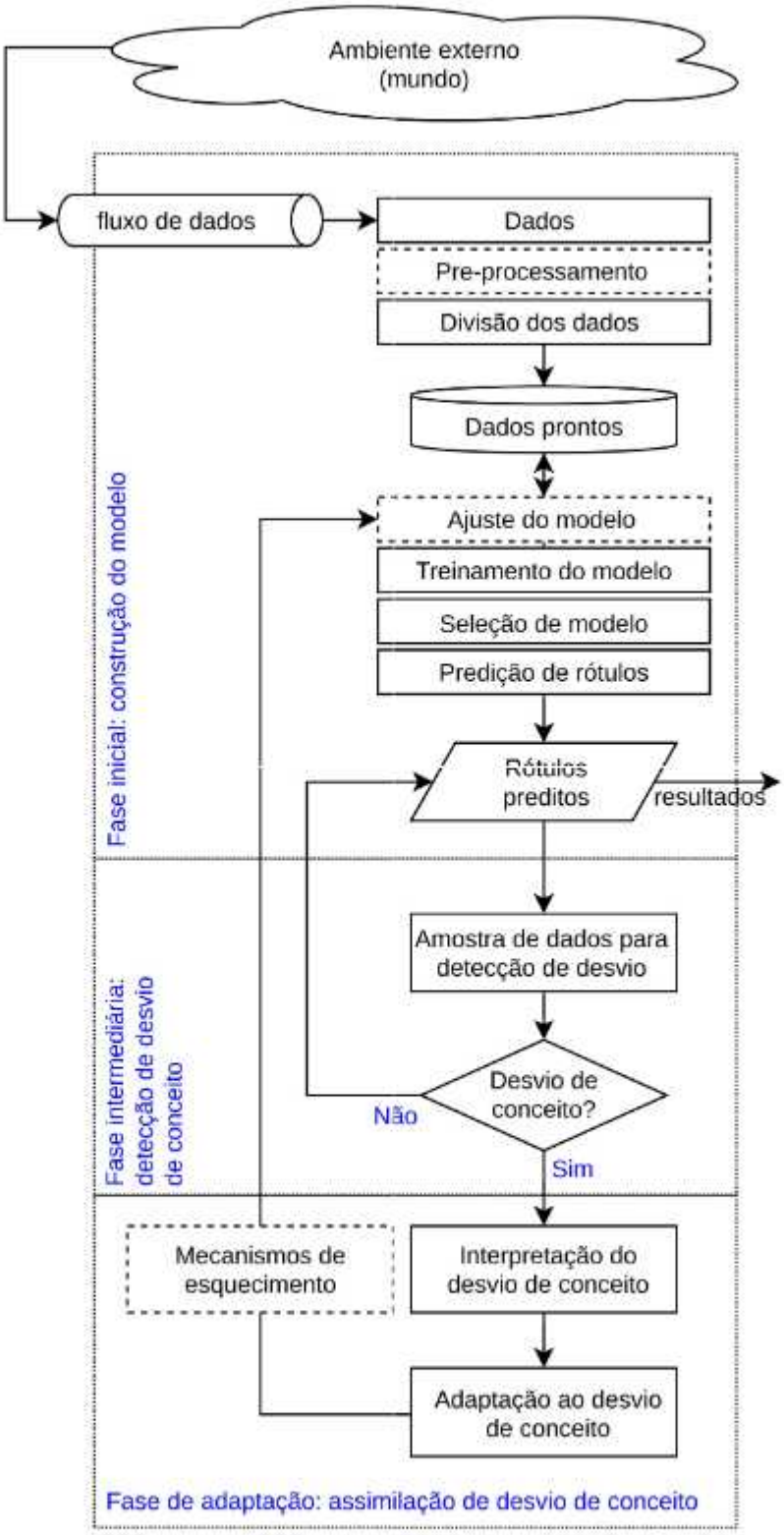


Figura 3.1: Fases do processamento de desvio de conceito (adaptado de Agrahari & Singh, 2021)

- Método baseado em janela, acumula as instâncias de dados de entrada a fim de formar um lote de dados (ou uma janela). Normalmente, nesse método são utilizadas duas janelas. A primeira janela é usada para armazenar os itens antigos enquanto a outra armazena os novos itens que estão chegando pelo fluxo de dados. A comparação entre os itens de cada janela permitirá identificar o desvio. O tamanho da janela pode ser fixo ou adaptável. Uma janela fixa significa que tem o mesmo tamanho de janela para toda a análise. Enquanto a janela adaptativa ajusta o tamanho com base nas condições de desvio. A janela é reduzida quando o desvio é detectado e expandida quando não há condição de desvio. Algumas técnicas conhecidas são: *Adaptive Windowing (ADWIN)* [106], ADWIN2 [107], SEED [108], *One Class Drift Detector (OCDD)* [109], *Ensemble Decision Trees for Concept drift (EDTC)* [110], *Statistical Test of Equal Proportions (STEPD)* [102], *Fast Hoeffding Drift Detection Method (FHDDM)* [111], *Wilcoxonr rank Sum Test Drift detector (WSTD)* [112], *Plover algorithm* [105] e *Paired Learner* [113].

- Método de análise de significância, utiliza a análise de hipóteses para detecção de desvio, geralmente é utilizado com o método baseado em janela e o método estatístico. Algumas técnicas conhecidas são: *Wilcoxonr rank Sum Test Drift detector (WSTD)* [112], *Fourier Inspired Windows for Concept Drift (FIWCD)* [19], *Discrete Fourier Transform (DFT)* [19], *Cumulative Sum (CUSUM) chart* [114], *Drift Detection Ensemble (DDE)* [115], *Hierarchical Hypothesis Testing with Classification Uncertainty (HHT-CU)* [116], *Two-Stage Multivariate Shift-Detection test based on EWMA (TSMSD-EWMA)* [117], *Hierarchical Linear Four Rates (HLFR)* [118], *Hierarchical Hypothesis Testing with Attribute-wise "Goodness-of-fit" (HHT-AG)* [116] e *Least Squares Density Difference (LSDD)* [92]

- Método baseado em distribuição de dados, utiliza dados históricos e itens de dados atuais para encontrar a mudança no contexto. Esse tipo de método, geralmente, é usado com a abordagem baseada em janela e analisa a significância estatística. O cálculo da mudança de distribuição de dados fornece informações sobre o local onde o desvio ocorreu. Entretanto, essa abordagem costuma apresentar custos computacionais consideráveis. Algumas técnicas conhecidas são: *Online Sequential Extreme Learning Machine (OS-ELM)* [119], *Self-Training Data Streams (STDS)* [120], *Principal Component Analysis based on Change Detection (PCA-CD)* [121], *Statistical Change Detection (SCD)* [122], *Least Squares Density Difference based on Change Detection Test (LSDD-CDT)* [123], *Local Drift Degree based Density Synchronized Drift Adaptation (LDD-DSDA)* [124], *Equal Density Estimation (EDE)* [125], *Least Squares Density Difference based on Change Detection Test (LSDD-CDT)* [123],

Competence Model (CM) [126] e *Hybrid Forest* [127]

- Método baseado em fronteiras de decisão, geralmente formam um limite usando itens inicialmente recebidos do fluxo de dados. A mudança na fronteira de decisão é considerada como um desvio. Algumas técnicas conhecidas são: *Margin Density Drift Detection (MD3)* [128], *Nearest Neighbor based on Density Variation Identification method (NN-DVI)* [129], *Self-adaption Neighborhood Density Clustering method (SNDC)* [130] e *Revising Density-based Spatial Clustering of Applications with Noise (Re-DBSCAN)* [131]

- Método dependente de modelo, cria um modelo de aprendizado usando algumas instâncias do fluxo de dados e verifica o desempenho do modelo para instâncias de fluxo de dados de entrada. Com a chegada de novos itens calcula a mudança na distribuição dos atributos do item recebido e, assim, tende a mudar a distribuição de probabilidade condicional $P(Classe\ Alvo|Atributos\ de\ Entrada)$. Esse tipo de método utiliza o teste de Kolmogorov-Smirnov, divergência KS [132], teste de soma de postos de Wilcoxon [133], etc. Algumas técnicas conhecidas são: ExStream [91] e um novo algoritmo de estabilidade para fluxo de dados não supervisionado [134].

- Método de análise sequencial, examina os itens de dados sequencialmente para identificar a mudança no contexto do fluxo de dados. Ele sinaliza o desvio quando a mudança na distribuição de dados excede o limite especificado. Tendo como inconveniente a necessidade de uma grande quantidade de exemplos de dados do novo conceito. Algumas técnicas conhecidas são: *Page-Hinkley Test (PHT)* [114, 135] *Forgetting Parameters Extreme Learning Machine (FP-ELM)* [94], *Online Sequential Extreme Learning Machine (OS-ELM)* [136] e *Diversity Measure as a new Drift Detection Method (DMDDM)* [137]

Um outro momento relevante nesse processo ocorre na terceira e última fase (adaptação), quando o desvio de conceito é identificado e o sistema atualiza o modelo preditivo. Isso é necessário porque os dados em movimento podem mudar ao longo do tempo. Geralmente, a combinação de métodos é o mais adequado para adaptação, pois combina vários modelos e suas previsões para prever a categoria dos novos itens recebidos pelo fluxo de dados [138]. Alguns fatores que fazem com que seja necessário um mecanismo de adaptação, são: nível de emprego da economia, mudança de idade, impacto sazonal, mudança de comportamento do mercado, mudança de demanda e oferta, mudança no arcabouço legal, eventos naturais extremos, etc.

A adaptação pode ocorrer de duas maneiras distintas. Primeiramente, de *forma implícita* ou passiva, onde o modelo preditivo é atualizado regularmente independente da detecção de um desvio de conceito, normalmente, a atualização ocorre em um intervalo de

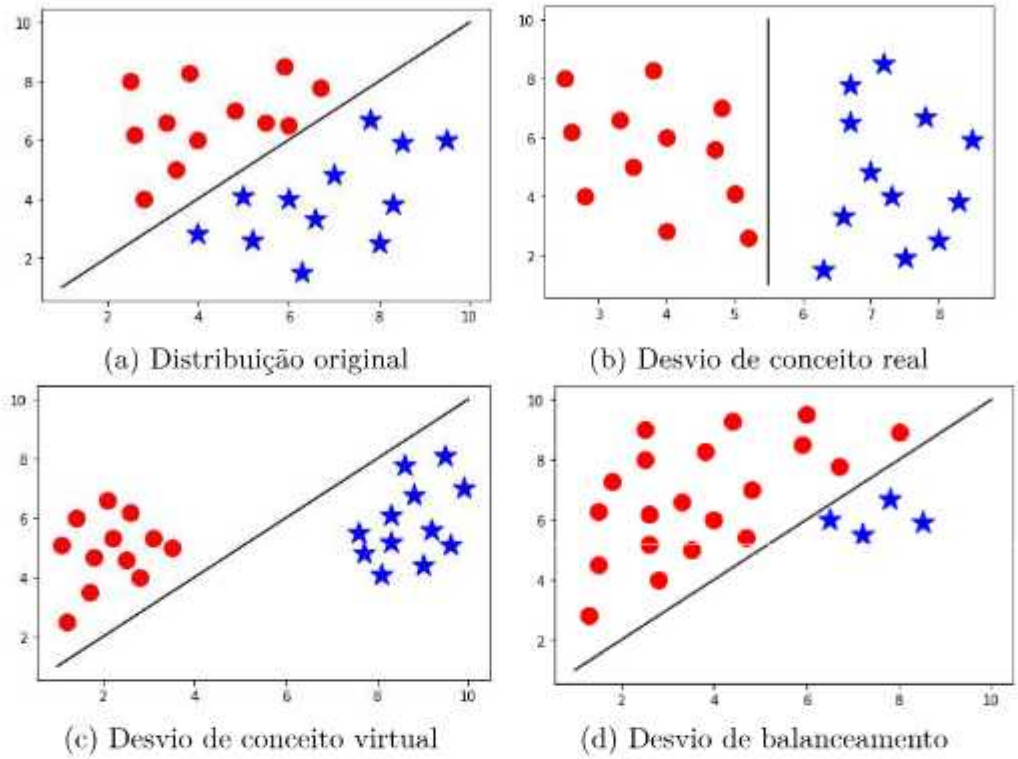
tempo específico. Porém, há um desperdício de recurso computacional, pois o modelo será atualizado mesmo sem que tenha ocorrido desvio de conceito [20]. A outra alternativa é o método explícito ou abordagem ativa, quando é feito o monitoramento de dados e, através de testes estatísticos, após identificar uma mudança nos dados o modelo é retreinado [19].

3.3.2 Tipificação de desvios de conceitos

A tipificação de desvio de conceito considera a fronteira de classificação, podendo ser um desvio de conceito virtual ou real. Em um desvio de conceito virtual observa-se que não ocorre um desvio na fronteira de classificação (decisão) apesar de uma mudança na distribuição dos dados, enquanto que em um desvio de conceito real a fronteira de classificação (decisão) muda à medida que a distribuição dos dados muda com o passar do tempo [19]. Na Figura 3.2 é possível visualizar essa definição. Onde a Figura 3.2a refere-se a distribuição original dos dados; a Figura 3.2b refere-se a representação de um desvio de conceito real ou mudança na fronteira de separação das classes, isto é, uma alteração na probabilidade *a posteriori* $P(y_k|x_i^{t+1})$; a Figura 3.2c refere-se a representação de um desvio de conceito virtual, isto é, uma alteração na probabilidade condicional $P(x_i^{t+1}|y_k)$, sem mudança na fronteira de separação das classes; e a Figura 3.2d refere-se a representação de uma alteração na proporção dos elementos de cada classe, isto é, uma alteração na probabilidade anterior da classe $P(y_k)$.

Em fluxo de dados, o tempo é um fator que deve ser considerado. Por exemplo, duas categorias distintas C1 e C2 podem ser classificadas, segundo a velocidade em que ocorre o desvio de conceito, da seguinte forma [139] (a Figura 3.3 ilustra esse exemplo):

- Desvio abrupto (*abrupt drift*): mudança repentina ou instantânea de conceito (Figura 3.3a).
- Desvio incremental (*incremental drift*): mudança gradual de conceito, sem oscilação, no decorrer do tempo (Figura 3.3b).
- Desvio gradual (*gradual drift*): mudança gradual, semelhante a uma mudança recorrente, entretanto nesse casos, o sistema demonstra uma estabilização em um novo conceito, que vai tornando-se cada vez mais frequente (Figura 3.3c).
- Desvio recorrente (*recurrent drift*): quando um conceito desaparece temporariamente e reaparecem após algum tempo (Figura 3.3d). Trata-se de um comportamento cíclico ou acíclico. Por exemplo, cíclico quando no inverno é possível observar um aumento nas vendas de roupas de frio e acíclico quando, devido a guerra, o preço do petróleo sofre elevação, voltando ao normal (preço anterior) em estado de paz.



(a) Distribuição original

(b) Desvio de conceito real

(c) Desvio de conceito virtual

(d) Desvio de balanceamento

Figura 3.2: Exemplo de dois modelos treinados a partir dados desbalanceados.

- Desvio pontual (*blip drift*): mudança de conceito pontual, com pouca ou pouquíssima frequência (Figura 3.3e).
- Ruído de desvio (*noise drift*): mudanças aleatórias que devem ser filtradas (Figura 3.3f).

Por fim, a técnica mais comum para lidar com desvio de conceito é a seleção de elementos mais atuais, usando-os na construção do modelo preditivo [44]. A ideia mais básica para lidar com o desvio de conceito usando a seleção de elementos são as janelas em movimento (*sliding window*), onde o modelo é reconstruído a cada movimento da janela [44].

Como os dados tendem a evoluir ao longo do tempo ocasionando alteração na distribuição dos dados, a relação entre uma observação e o rótulo $y = h(X)$ pode mudar. Portanto, o algoritmo de atualização do modelo f precisa incluir alguns mecanismos de esquecimento para que o modelo possa se adaptar à nova distribuição de dados à medida que recebe novas informações [140]. Nesse sentido, alguns algoritmos já foram propostos: um algoritmo de esquecimento gradual que considerava a idade da amostra [141] e dois métodos de ponderação de tempo: (i) um esquema global e (ii) um esquema local, baseados em uma função exponencial [142].

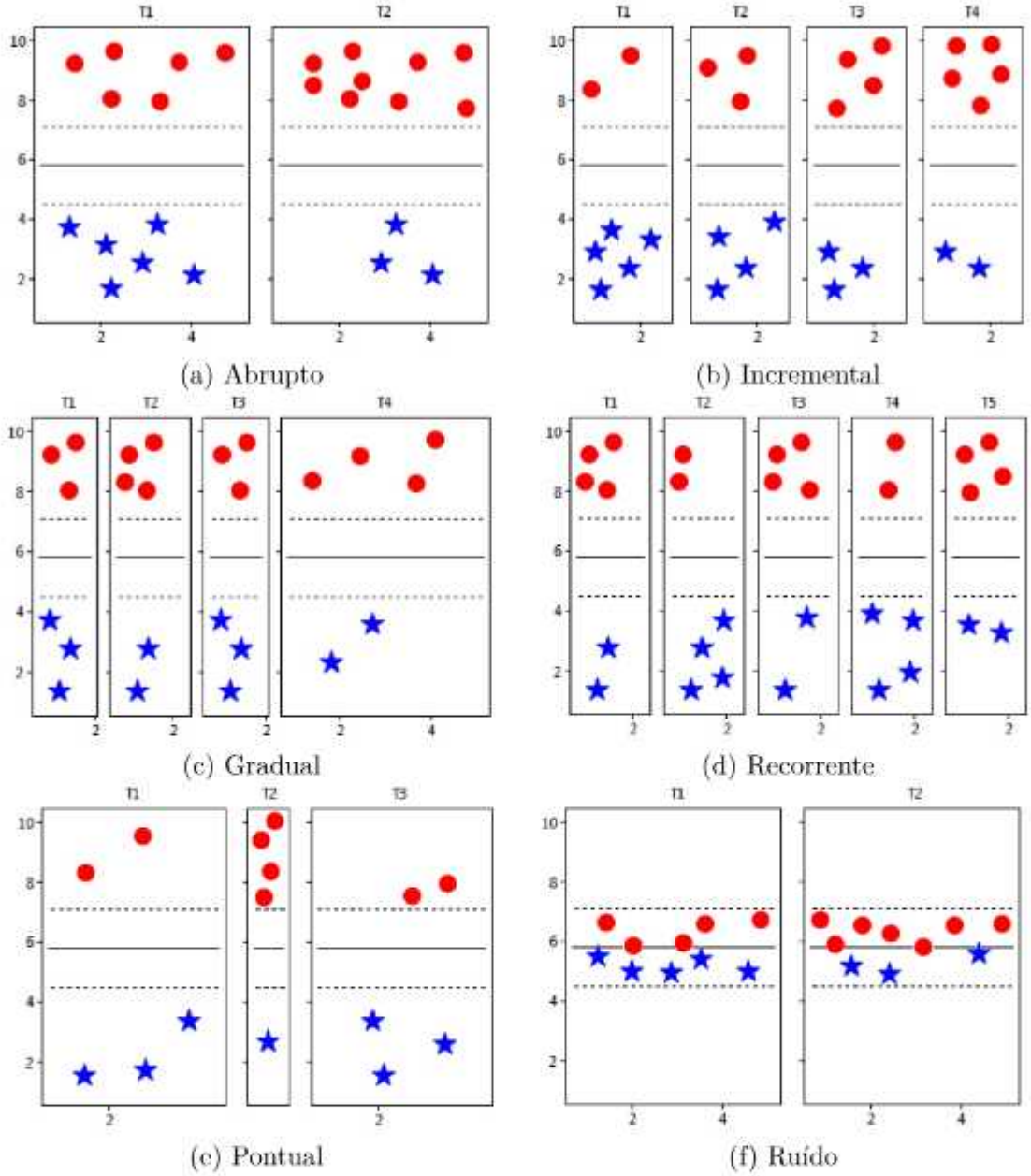


Figura 3.3: Classificação de desvios de conceito quanto ao tempo.

Este estudo concentra-se no desvio de conceito relacionado a probabilidade anterior da classe, $P(y_k)$. Isto é, o desbalanceamento dos dados recebidos. A seção seguinte aborda essa questão com mais detalhes e apresenta algumas técnicas que podem ser utilizadas para minimizar esse problema.

3.4 Desbalanceamento

O desbalanceamento em um conjunto de dados acontece quando as categorias de dados não são representadas igualmente, ou seja, pelo menos uma categoria é minoritária em relação às outras categorias [26]. Isso pode causar viés de aprendizagem para a categoria majoritária e prejudicar a generalização do modelo. Existe dois tipos de desbalanceamento, o intrínseco, quando o desbalanceamento é algo natural do problema, por exemplo, a situação financeira das empresas que normalmente são saudáveis, com uma minoria em DF. O outro tipo de desbalanceamento, o extrínseco, ocorre quando o desbalanceamento decorre de um falha na coleta dos dados [1]. Aprender com conjuntos de dados desequilibrados é um desafio, pois os algoritmos padrões de AM são projetados para otimizar a generalização e, como consequência, a categoria minoritária pode ser completamente ignorada [26].

O problema de aprendizado torna-se particularmente desafiador quando ele é afetado por desvio de conceito e desbalanceamento, prejudicando significativamente o desempenho preditivo do modelo [19]. Esse desafio é porque o desequilíbrio de classes pode afetar o tratamento de desvio de conceito [26]. Por exemplo, algoritmos de detecção de desvio baseados no erro de classificação tradicional podem ser sensíveis ao grau de desbalanceamento e tornam-se menos eficazes, e as técnicas de desbalanceamento de categoria precisam ser adaptáveis às taxas de desbalanceamento variáveis; caso contrário, a categoria que recebe o tratamento preferencial pode não ser a categoria minoritária correta naquele momento.

O tratamento do desequilíbrio de classes pode ser feito de três formas, abordagem de pré-processamento, abordagem de aprendizado em algoritmo e abordagem híbrida ou aprendizado em comitê. A abordagem de aprendizado em algoritmo tem sua estratégia baseada em definir diferentes pesos para predições incorretas, provocando um crescimento no custo do erro da categoria minoritária, o que ocasiona um viés no processo de aprendizado a seu favor. A abordagem de pré-processamento baseia-se em duas estratégias: remoção de elementos da categoria majoritária (*undersampling*) e adição de elementos sintéticos no conjunto de categorias minoritárias (*oversampling*); entretanto, devido a necessidade de calcular múltiplas distâncias entre os elementos utilizados no treinamento essas estratégias tendem a ser custosas, mesmo em processamentos em *batch*, *i.e.* SMOTE [143]. Na abordagem híbrida ocorre um aprendizado conjunto, onde é utilizado tanto a abodagem de pré-processamento como a abordagem de aprendizado em algoritmo [144].

O contexto de fluxo de dados acrescenta outros desafios à tarefa de tratar com dados desbalanceados. Por exemplo, em uma distribuição conhecidamente balanceada, em um período determinado de tempo, uma de suas categorias poderia ser sub-representada, ocasionando um desbalanceamento temporário. Um outro complicador seria uma alteração

na distribuição de probabilidade da categoria, o que seria um desvio de conceito ou uma uma evolução de conceito, que ocorre quando uma das categorias deixa de existir. Estas dificuldades motivaram o surgimento de alguns métodos para tratar do desbalanceamento em fluxo de dados, já listados na Tabela 2.4.

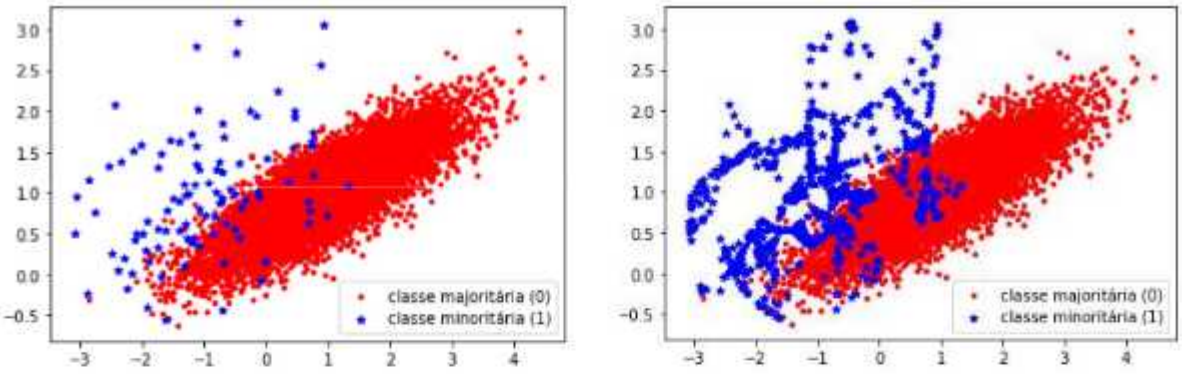
Além disso, algumas métricas tradicionalmente utilizadas para avaliar modelos de AM (*i.e.* acurácia) não são adequadas para dados desbalanceados [44]. Isso ocorre quando a métrica utiliza o número de elementos da classe majoritária e, devido ao desequilíbrio entre as classes, essa classe acaba tendo maior influência na métrica e enviesa resultado. Por exemplo, quando temos uma situação de desbalanceamento, onde 90% dos elementos pertencem à classe majoritária e 10% pertencem à classe minoritária. Uma acurácia de 90% não representa necessariamente um bom desempenho, pois o modelo poderia está acertando apenas elementos da classe majoritária e errando todos da classe minoritária. No caso de empresas em DF, é na identificação da classe minoritária que está o real valor para o negócio. Portanto, a acurácia não é uma medida segura.

3.4.1 SMOTE

SMOTE é um algoritmo que realiza aumento de dados criando pontos de instâncias sintéticas com base nos pontos de instâncias originais. O SMOTE pode ser visto como uma versão avançada de sobreamostragem ou como um algoritmo específico para aumento de dados. A vantagem do SMOTE é a não geração de duplicatas, mas sim criando pontos sintéticos ligeiramente diferentes dos pontos originais. Seu funcionamento pode ser resumido da seguinte forma [143]:

1. Selecionar um elemento aleatório m da classe minoritária;
2. Identificar os k vizinhos mais próximos;
3. Calcular os vetores (de atributos) entre m e cada k vizinho selecionado (distância entre cada atributo);
4. Multiplicar cada vetor por um número aleatório entre 0 e 1;
5. Adicionar cada vetor gerado ao ponto de dados atual m e gerar uma nova instância (sintética).

Esta operação assemelha-se ao movimento do elemento m na direção de um vizinho. Dessa forma, você garante que seu ponto de dados sintéticos não seja uma cópia exata de um ponto de dados existente, ao mesmo tempo em que não é muito diferente das observações conhecidas em sua classe minoritária.



(a) Desbalanceado (b) Balanceado com SMOTE

Figura 3.4: Comparativo de amostras antes (a) e após aplicação da técnica de SMOTE (b)

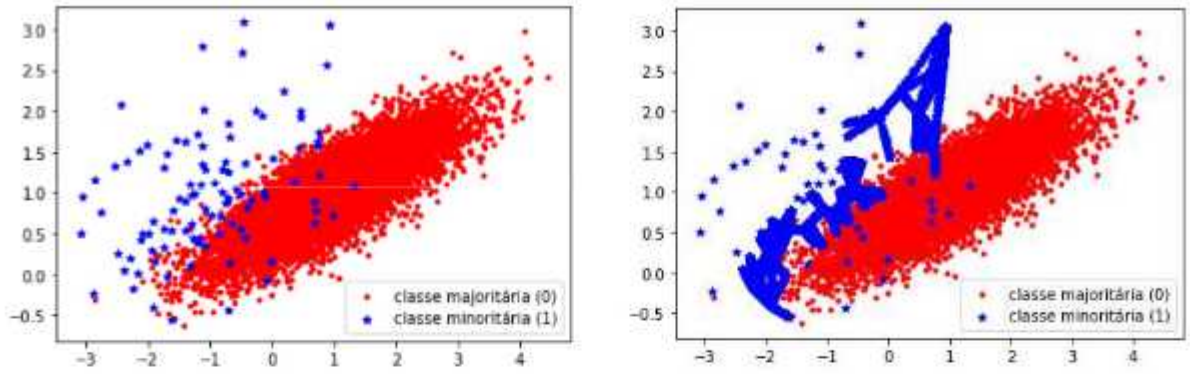
A aplicação desse método pode ser observada em uma amostra binária de 1000 elementos, ilustrado na Figura 3.4. Na Figura 3.4a, onde 0 é a classe majoritária com 990 elementos (em vermelho) e 1 é a classe minoritária com 10 elementos (em azul) na proporção de 1:100 (aproximadamente), ambas com dois atributos. Na Figura 3.4b, após execução desse método foi obtida uma amostra com a mesma quantidade de elementos 0 (vermelho) e 1 (azul), proporção de 1:1, totalizando 1.980 elementos. É possível observar que os elementos sintéticos foram gerados respeitando a vizinhança dos elementos reais.

Entretanto, essa técnica tem algumas desvantagens, uma delas a influência negativa de uma grande quantidade de exemplos sintéticos [143]. Uma forma de tentar minimizar esse problema é por meio da seleção de exemplos da classe minoritária para serem superamostrados usando o SMOTE. O método *Borderline-SMOTE* (B-SMOTE) é um exemplo de variante do SMOTE.

3.4.2 Borderline-SMOTE

Uma extensão popular para SMOTE, o B-SMOTE, envolve selecionar com um modelo de classificação de k-vizinhos (*K-Nearest Neighbor*, KNN) as instâncias da classe minoritária consideradas como mais difíceis de classificar. Após a seleção das instâncias é aplicada a superamostragem (SMOTE), fornecendo instâncias sintéticas em regiões mais relevantes para diferenciação das classes [145].

Esses exemplos que são classificados incorretamente são provavelmente ambíguos e estão em uma região da borda ou fronteira do limite de decisão onde a associação de classe pode se sobrepor. Como tal, essa modificação no SMOTE é denominado B-SMOTE e foi proposto por Han, H. *et al.* (2005) [145]. Os autores também descrevem uma versão do método que superamostra a classe majoritária para aqueles exemplos que causam uma classificação incorreta de instâncias limítrofes na classe minoritária. Isso é chamado de



(a) Desbalanceado (b) Balanceado com B-SMOTE

Figura 3.5: Comparativo de amostras antes (a) e após aplicação da técnica de B-SMOTE (b)

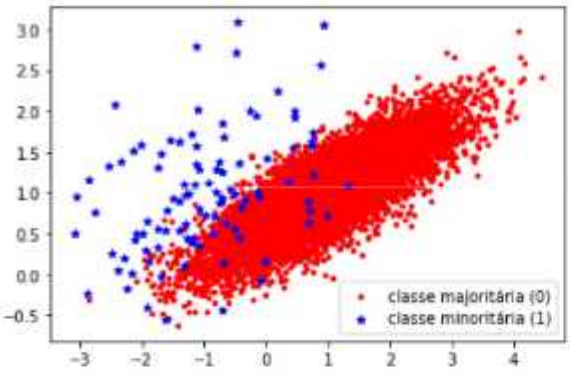
Borderline-SMOTE1, enquanto a sobreamostragem apenas dos casos limítrofes na classe minoritária é chamada de Borderline-SMOTE2 [145].

A aplicação desse método pode ser observada em uma amostra binária de 1000 elementos, ilustrado na Figura 3.5. Na Figura 3.5a, 0 é a classe majoritária com 990 elementos (em vermelho) e 1 é a classe minoritária com 10 elementos (em azul) na proporção de 100:1 (aproximadamente), ambos com dois atributos. Na Figura 3.5b, após execução desse método foi obtida uma amostra com a mesma quantidade de elementos 0 (vermelho) e 1 (azul), proporção de 1:1, totalizando 1.980 elementos. É possível observar que não houve geração de elementos sintéticos próximos aos pontos mais afastados (à direita) da fronteira das classes.

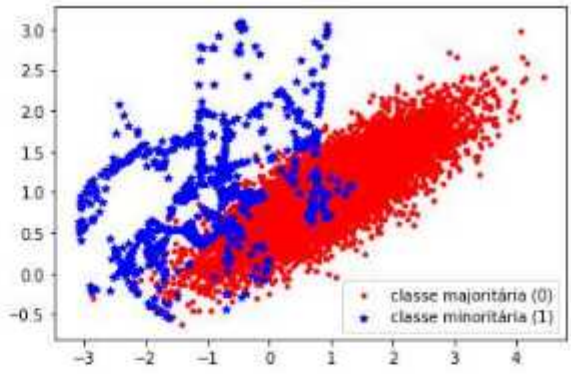
3.4.3 Sub-amostragem

Outra forma de amenizar a influência negativa de uma grande quantidade de exemplos sintéticos na classe minoritária, após o balanceamento, é aplicar uma técnica de subamostragem na classe majoritária [143]. Dessa forma, é possível aumentar a proporção de elementos reais em relação aos sintéticos. A aplicação da subamostragem juntamente com a sobre amostragem tem demonstrado resultado melhores que a aplicação isolada das ténicas [143].

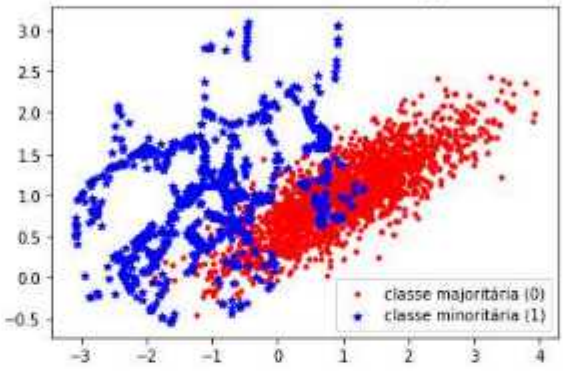
A aplicação desse método pode ser observada na Figura 3.6. Na Figura 3.6a, uma amostra binária de 1000 elementos, onde 0 é a classe majoritária com 990 elementos (em vermelho) e 1 é a classe minoritária com 10 elementos (em azul) na proporção de 1:100 (aproximadamente), ambas com dois atributos. Na Figura 3.6b, é possível observar que novos elementos foram adicionados para atingir a proporção de 1:10. Na Figura 3.6c, com a aplicação da subamostragem da classe majoritária, são removidos alguns elementos fazendo com que a proporção passe a ser 1:2.



(a) Desbalanceado



(b) Balanceado com SMOTE



(c) Balanceado com SMOTE e subamostragem

Figura 3.6: Aplicação de técnicas de reamostragem (b) e subamostragem (c).

3.5 Trabalhos relacionados

A busca por uma forma de predizer falência ou identificar o momento de DFs de empresas não é algo recente. Barboza *et al.* (2017) [7] reconhece essa busca por parte do mercado e de acadêmicos. Apresenta como métodos tradicionais (*e.g.* ADM e RL) aqueles que fazem uso de técnicas estatísticas como em Altman (1968) [3], que evidencia décadas de interesse nessa área. Eles também destacam que as técnicas de AM apresentam, em média, uma acurácia 10% superior que as técnicas tradicionais. Esse resultado foi obtido a partir de experimentos com dados de empresas da América do Norte, entre os anos de 1985 a 2013, de onde foram selecionados 11 atributos, e de modelos preditivos que utilizam técnicas de: *Bagging*, *Boosting*, RF, SVM, Redes Neurais Artificiais (RNA), RL e ADM. Por fim, sugerem a utilização de outras métricas além da acurácia e novas variáveis para os modelos.

Em seguida, Alaka *et al.* (2017) [8] publicaram uma revisão sistemática, baseada em artigos publicados entre 2010 e 2015, sobre predição de falência abordando o uso de técnicas estatísticas: ADM e RL; e seis técnicas de IA: RNA, SVM, *rough sets*, RBC, AD e AG. No geral, verificou-se que nenhuma técnica isolada é predominantemente melhor do que outras técnicas. Conclui-se que um modelo de desempenho global melhor só pode

ser encontrado pela integração informada de técnicas para formar um modelo híbrido. Com isso, o artigo contribuiu para um entendimento aprofundado das características das ferramentas utilizadas para desenvolver modelos de previsão de falência e suas deficiências. Termina por sugerir a utilização de técnicas mais sofisticadas como XGBoost e algumas outras.

Clement, C. (2020) [31] complementa a revisão anterior com informações de 32 artigos, publicados entre os anos de 2016 a 2020, tratando de predição de falência e AM. Entre os artigos estudados foi observada uma variação de acurácia entre 70% e 99,2% com a utilização de diferentes modelos preditivos. Desde técnicas paramétricas: ADM, RL, Análise Discriminante Linear (ADL), ADM, a técnicas não paramétricas: AdaBoost, RBC, *Extreme learning machine*, *fuzzy-set*, RNA, *Gaussian processes*, AD, *Decision rule inducer*, KNN, XGBoost, RF, entre outros. Por fim, conclui que não existe uma técnica que seja melhor e que a precisão dos resultados depende mais do ajuste do algoritmo com base na amostra utilizada e suas propriedades, em linha com o que já havia sido mencionado por Alaka *et al.* (2017) [8].

Ao abordar a questão da predição (identificação) de DF, Barboza *et al.* (2021) [40] utiliza técnicas de AM (*i.e.* XGBoost e RF) atingindo acurácia de 96%. Nesse estudo foram utilizados 17 atributos extraídos de dados de empresas de países da América Latina, incluindo o Brasil, entre os anos de 2000 a 2017. Devido ao forte desbalanceamento de dados, a técnica de SMOTE foi utilizada para diminuir o desequilíbrio entre as classes e seus efeitos no desempenho do modelo. Nesse contexto, além da acurácia, foram utilizadas métricas adequadas à avaliação dos resultados do modelo, como: AUC-ROC, tipo de erro I e II. Apesar do bom resultado obtido, foi observado que os modelos utilizados são de difícil interpretação e as variáveis de entrada permanecem uma questão em aberto, devendo ser explorada em novos estudos.

Sun, J. *et al.* (2017) [42] reconhece a carência de estudos capazes de predizer a DF de maneira dinâmica com dados não estacionários. Nesse estudo foram utilizados 7 atributos extraídos de 932 empresas chinesas listadas em bolsa, entre os anos de 2005 e 2012. O desvio de conceito foi tratado através de um mecanismo de esquecimento baseado em pesos [142]. Por meio de modelos baseados em SVM foi possível obter até 87,60% de acurácia. Entretanto, apesar de ser um problema comum, em situação de identificação de DF, a questão do desbalanceamento de classes não foi abordada neste estudo.

Complementando o estudo anterior, Shen, F. *et al.* (2020) [44] trata a questão de predição de DF em um contexto de dados não estacionários, considerando a questão do desbalanceamento de classes presente na base de dados de empresas chinesas listadas, entre os anos de 2007 a 2017, de onde foram extraídos 70 atributos. O desvio de conceito foi tratado através de um mecanismo de esquecimento baseado em pesos [142] e um meca-

nismo de janelas deslizantes. A junção das técnicas de ANS e REA permitiu propor uma nova técnica (ANS-REA), que foi utilizada para tratar o desequilíbrio entre as classes. Os autores destacam a importância de utilizar métricas adequadas ao contexto de desbalanceamento de classes. Assim, o RF associado a técnica de balanceamento ANS-REA apresentou os melhores resultados, obtendo 91% de acurácia média (AUC-ROC), índice *Kappa* médio de 71%, F_1 médio de 80% e G_{mean} médio de 87%.

É possível observar que a predição de DF vem evoluindo e novas camadas de conhecimento são adicionadas com o tempo. No momento atual busca-se identificar a DF através de dados não estacionários quando também existe um desequilíbrio de classes. Formas de medir os resultados produzidos pelos modelos preditivos é essencial para guiar os experimentos. Saito, T. & Rehmsmeier, M. (2015) [146] demonstra que a interpretabilidade visual dos gráficos ROC no contexto de conjuntos de dados desequilibrados pode ser enganosa em relação às conclusões sobre a confiabilidade do desempenho da classificação, devido a uma interpretação intuitiva, mas errada da especificidade. Por outro lado, os gráficos das Curva de Precisão e Sensibilidade (PS) podem fornecer ao leitor uma previsão precisa do desempenho futuro da classificação devido ao fato de avaliarem a fração de verdadeiros positivos entre as previsões positivas.

Apesar da quantidade de estudos sobre a questão de DF é possível observar uma concentração de estudos nos EUA e em países que tiveram os dados de empresas disponibilizados publicamente, em UCI, como Taiwan [41, 53] e Polônia [38, 69] (Figura 2.3). No Brasil, dados de empresas listadas na bolsa (B3) são publicados no portal de dados abertos da CVM, entretanto não estão prontos para processamento, necessitando de esforço para obter os indicadores econômico-financeiros.

Por fim, este trabalho procura utilizar várias técnicas encontradas na literatura a fim de identificar situações de DF de empresas em ambientes de fluxo de dados com desequilíbrio entre classes. Aplicando os modelos preditivos com melhor desempenho nos estudos mais recentes e utilizando dados de empresas brasileiras, que até então não tem sido utilizados em estudos dessa natureza. Por meio de uma solução semelhante a adotada por Shen, F. *et al.* (2020) [44] e métricas adequadas para avaliação de desempenho de modelos com dados desbalanceados pretende-se contribuir com uma solução capaz de predizer a DF em corporações no Brasil.

3.6 Métricas de avaliação

A qualidade dos algoritmos de aprendizado geralmente é avaliada analisando o desempenho deles nos dados de teste [147]. Para tanto, as previsões geradas pelos modelos são comparadas com as verdadeiras classes de dados do conjunto de teste (que estão ocultas

Tabela 3.1: Matriz de confusão

Matriz de Confusão		Classe predita		
		Positivo	Negativo	Total
Classe real	Positivo	VP	FN	**P**
	Negativo	FP	VN	**N**
	Total	**PP**	**NP**	**T**

dos modelos para fins de avaliação). A partir desse resultado são calculadas algumas medidas de desempenho. Uma maneira de resumir o desempenho dos modelos é fazer uma tabulação cruzada entre as classes reais e previstas. A tabulação cruzada resultante é uma matriz, chamada matriz de confusão. As colunas da matriz de confusão representam as contagens de instâncias nas classes previstas, enquanto as linhas representam as contagens de instâncias nas classes reais (ou vice-versa) [26].

Na Tabela 3.1 pode ser observada uma matriz de confusão. Ao cruzar os valores preditos (colunas) e os valores reais (linhas) encontram-se os valores de Verdadeiro Positivo (VP), Falso Negativo (FN), Falso Positivo (FP) e Verdadeiro Negativo (VN), com o somatório representado por T (VP + FN + FP + VN). Ao somar cada coluna encontram-se os valores Total de Positivos Preditos (PP) e Total de Negativos Preditos (NP), (VP + FP) e (FN + VN), respectivamente. Ao somar cada linha encontram-se os valores Total de Positivos Reais (P) e Total de Negativos Reais (N), (VP + FN) e (FP + VN), respectivamente.

Os algoritmos padrões de AM geralmente são tendenciosos para a classe majoritária, uma vez que as regras que predizem corretamente essas instâncias são ponderadas positivamente em favor da métrica de precisão ou da função de custo correspondente [26]. Por outro lado, regras específicas que preveem exemplos da classe minoritária podem ser ignoradas (tratando-as como ruído), uma vez que regras mais gerais são preferidas. Como consequência, as instâncias da classe minoritária são mais frequentemente mal classificadas do que as da majoritária [26].

Conclui-se que a acurácia não é uma medida adequada no cenário de desequilíbrio, pois não faz distinção entre o número de exemplos corretamente classificados de diferentes classes [26]. No contexto desse estudo, a classe minoritária ou classe positiva são as empresas em DF. Modelos que apresentam uma alta acurácia global e não tem capacidade de distinguir a classe minoritária (empresas em DF) não têm efeito prático [26]. Portanto, medidas mais informativas são necessárias para avaliar a qualidade dos modelos, por exemplo, AUC [24, 44, 57, 64], média geométrica [44], $F_{measure}$ [44, 57], precisão e sensibilidade [24, 57, 60, 64].

3.6.1 Precisão

A precisão avalia a fração de instâncias classificadas corretamente entre as classificadas como positivas [26]. Neste caso, é a fração de instâncias corretamente classificadas como em DF (VP) em relação ao total de instâncias classificadas em DF (VP + FP).

$$precisao = \frac{VP}{PP} = \frac{VP}{VP + FP}$$

3.6.2 Revocação

É a fração do total de instâncias positivas classificadas corretamente, isto é Taxa de Verdadeiro Positvo (TVP) [26], tratuzida do inglês *recall*, também conhecida como sensibilidade. Neste estudo, são as instâncias corretamente classificadas em DF (VP) em relação às instâncias realmente positivas (VP + FN).

$$recall = \frac{VP}{P} = \frac{VP}{VP + FN}$$

3.6.3 Média harmônica

$F_{measure}$ é uma medida que foca na análise de classes positivas. Seu objetivo é analisar o equilíbrio entre correção e cobertura na classificação de instâncias positivas. Para tanto, a medida utiliza uma média harmônica ponderada entre precisão e revocação (sensibilidade) [26]. O parâmetro β controla a importância dada a cada termo. Somente quando os valores de sensibilidade e precisão são grandes o suficiente, a $F_{measure}$ pode atingir um valor alto, onde o valor máximo é 1.

$$F_\beta = \frac{(1 + \beta^2) \times precisao \times recall}{\beta^2 \times precisao + recall}$$

Uma escolha comum é definir $\beta = 1$, levando à medida F_1.

$$F_1 = \frac{2 \times precisao \times recall}{precisao + sensibilidade}$$

3.6.4 Média geométrica

Uma variação do $F_{measure}$ é o $G_{measure}$, que usa a média geométrica em vez da média harmônica para compensar a relação de precisão e sensibilidade. Somente quando a precisão e a sensibilidade são grandes o suficiente, a $G_{measure}$ pode atingir um valor alto, onde o valor máximo é 1.

$$G_{measure} = \sqrt{precisao \times sensibilidade}$$

G_{mean} também faz uso de média geométrica, mas utiliza informações de ambas as classes, pois é a média geométrica do produto entre sensibilidade (TVP) e especificidade (TVN). Esta medida visa um equilíbrio entre os desempenhos de classificação nas classes majoritária e minoritária. Um desempenho ruim na previsão dos exemplos positivos levará a um valor G_{mean} baixo, mesmo que os exemplos negativos sejam classificados corretamente. Quando o modelo tem desempenho igual em ambas as classes, ou quando as classes são igualmente balanceadas, essa medida é equivalente à acurácia convencional (Ac). No entanto, se a precisão convencional for alta apenas porque o modelo tira vantagem de um conjunto de teste desbalanceado, a precisão balanceada será menor [26].

$$G_{mean} = \sqrt{especificidade \times sensibilidade}$$

3.6.5 Curva ROC

A curva ROC é o gráfico de características operacionais do receptor (*Receiver Operating Curve*, abreviado como ROC) é uma técnica para visualizar, organizar e selecionar modelos com base em seu desempenho [148]. O gráfico ROC é representado pela Taxa de Falso Positivo (FP/N) no eixo x e pela Taxa de Verdadeiro Positivo (sensibilidade, VP/P) no eixo y, trançando uma curva com diferentes limiares de classificação [26]. Portanto, é uma representação bidimensional do desempenho do modelo. A comparação entre modelos pode ser feita reduzindo a curva ROC a um único valor escalar que representa o desempenho esperado [148]. O método mais comum é calcular a área sob a curva ROC, abreviada como AUC-ROC (*Area Under the Curve - ROC*) [149, 150]. Nessa curva, o valor da área estará sempre entre 0 e 1. Nenhum modelo realista deve ter uma AUC menor que 0,5 [148], pois esse é o valor de referencia da adivinhação aleatória, que pode ser presesentada pela diagonal do quadrado entre os pontos (0, 0) e (1, 1).

3.6.6 Curva de precisão e sensibilidade

Em algumas situações a AUC-ROC não é o indicador de desempenho ideal para avaliar um modelo. Quando os dados estão fortemente desbalanceados, esse indicador pode mascarar um baixo desempenho no reconhecimento de instâncias positivas (empresas em DF). Em bases de dados binárias, com desbalanceamento superior a 1:10, é possível verificar esse problema [146]. Assim, recomenda-se o uso da PS, como forma de complementar a análise.

Uma curva PS é simplesmente um gráfico com valores de precisão no eixo y e valores de sensibilidade no eixo x. Em outras palavras, a curva PS contém $\frac{TP}{(TP+FN)}$ no eixo y e $\frac{TP}{(TP+FP)}$ no eixo x. Tendo como referência a taxa de desbalanceamento do conjunto que marca o desempenho aleatório.

Capítulo 4

Metodologia

Neste capítulo é apresentada a metodologia proposta para o desenvolvimento da pesquisa, explicando como as hipóteses de pesquisa serão testadas através de experimentos que utilizam dados da CVM. São apresentados os *frameworks* de AM, algoritmos e medidas de avaliação utilizadas. A Seção 4.1 explica como os dados foram encontrados na CVM e o que foi preciso para extrair os indicadores. A Seção 4.2 descreve como a abordagem de fluxo de dados foi aplicada neste estudo para predição de DF. Por fim, a Seção 4.3 detalha o processo de avaliação dos modelos em um cenário de dados sequenciais e não estacionários.

4.1 Tratamento dos dados

Este estudo faz uso de dados disponibilizados pela CVM em seu portal de dados abertos[1], onde é possível encontrar a base de dados de informações cadastrais das companhias abertas[2] e fortemente desbalanceada em relação a DF.

A variável *situação emissor* contém informações relevantes para categorização das companhias e treinamento do modelo preditivo, sendo utilizada para identificar empresas em DF. Essa variável diz respeito a uma situação especial da companhia, podendo assumir os seguintes valores: "fase pré-operacional", "fase operacional", "liquidação extrajudicial", "em recuperação judicial ou equivalente", "falida", "em liquidação judicial", "em recuperação extrajudicial" e "paralisada". Entre esses valores, destacam-se a "fase pré-operacional", que deve ser desconsiderada uma vez que a companhia ainda não iniciou suas operações junto a CVM; a fase operacional, em que a empresa está operando e, portanto, podendo ser avaliada quanto a sua *situação financeira* (*i.e.* em DF); e, os demais valores que juntos caracterizam a DF.

[1] http://dados.cvm.gov.br/

[2] http://dados.cvm.gov.br/dataset/cia_aberta-cad

Tabela 4.1: Panorama do desbalanceamento da base de dados em um trimestre específico.

Categoria	**Quant.**	**Perc.**
Fase pré-operacional	57	7,26%
Fase operacional	684	87,13%
Paralisada	3	0,38%
Em recuperação judicial ou equivalente	28	3,57%
Liquidação extrajudicial	8	1,02%
Falida	2	0,25%
Em liquidação judicial	2	0,25%
Em recuperação extrajudicial	1	0,13%
Total	**785**	**100%**

Outra variável relevante é a *situação*, que pode assumir os valores[3]: "ativa", "cancelada", "suspensa" e "inadimplente".

- Quando "ativa" a companhia que está fornecendo informações à CVM no prazo estipulado, tal como definido no art. 21 da instrução CVM 480 [5];
- A situação "inadimplente" é anterior à suspensão do registro. A CVM publica a lista dos emissores de valores mobiliários em atraso de pelo menos três meses no cumprimento de suas obrigações periódicas, tal divulgação, que alerta o mercado e os investidores sobre a situação, ainda não afeta o registro da empresa;
- A "suspensão" ocorre caso a situação persista e o descumprimento ocorra por período superior a 12 meses, assim a CVM suspende o registro da companhia aberta. A partir daí, a empresa não pode ter os valores mobiliários negociados em mercados regulamentados (*i.e.* B3).
- O "cancelamento" ocorre se a suspensão do registro perdurar por período superior a 12 meses. Nesse caso a CVM cancela o registro da companhia aberta, o que também ocorre se a companhia for extinta.

Em novembro de 2021, o arquivo contendo os dados de informações cadastrais das companhias abertas contava com 2.485 registros, 785 em situação "ativo". Analisando apenas os registros ativos, observou-se um forte desbalanceamento da base de dados, como descrito em Tabela 4.1.

Neste estudo, desconsiderando a categoria fase pré-operacional, as categorias foram organizadas da seguinte forma:

- SDF (94%)

[3]https://www.gov.br/cvm/pt-br/assuntos/protecao/alertas/alertas-sobre-companhias

  - Fase operacional

- Em DF (6%)
  - Paralisada
  - Em recuperação judicial ou equivalente
  - Liquidação extrajudicial
  - Falida
  - Em liquidação judicial
  - Em recuperação extrajudicial

No Portal de Dados Abertos da CVM foram acessados os documentos periódicos e eventuais de empresas reguladas. O documento Formulário de Informações Trimestrais (ITR)[4] foi utilizado para calcular indicadores econômico-financeiros, em um período de 10 anos (2011 a 2020). Entre os arquivos deste formulário, foram utilizados:

- Balanço Patrimonial de Ativos (BPA)
- Balanço Patrimonial de Passivos (BPP)
- Demonstração de Resultado (DRE)
- Demonstração de Fluxo de Caixa (DFC)

Outro documento foi o Formulário de Referência (FRE)[5] de onde foram obtidas informações sobre a quantidade de ações negociadas na B3, contidas no arquivo de distribuição de capital. Os indicadores obtidos desses arquivos foram extraídos diretamente de colunas específicas (Apêndice A.1.2) e indiretamente, calculados a partir de diferentes valores em um mesmo arquivo ou em diferentes arquivos (Apêndice A.1.2). O resultando foi um conjunto de 84 indicadores (Apêndice A).

4.1.1 Valores ausentes

A depender do setor de atuação da empresa, algumas contas podem estar ausentes. Por exemplo, instituições financeiras, classificadas pela CVM no setor de atividade "bancos" ou "serviços financeiros", são empresas prestadoras de serviço. Elas não possuem valores na conta *Estoque*, pois não trabalham com produção, venda ou revenda de mercadorias. Portanto, nesses casos, as contas receberam o valor zero.

[4]http://dados.cvm.gov.br/dataset/cia_aberta-doc-itr
[5]http://dados.cvm.gov.br/dataset/cia_aberta-doc-fre

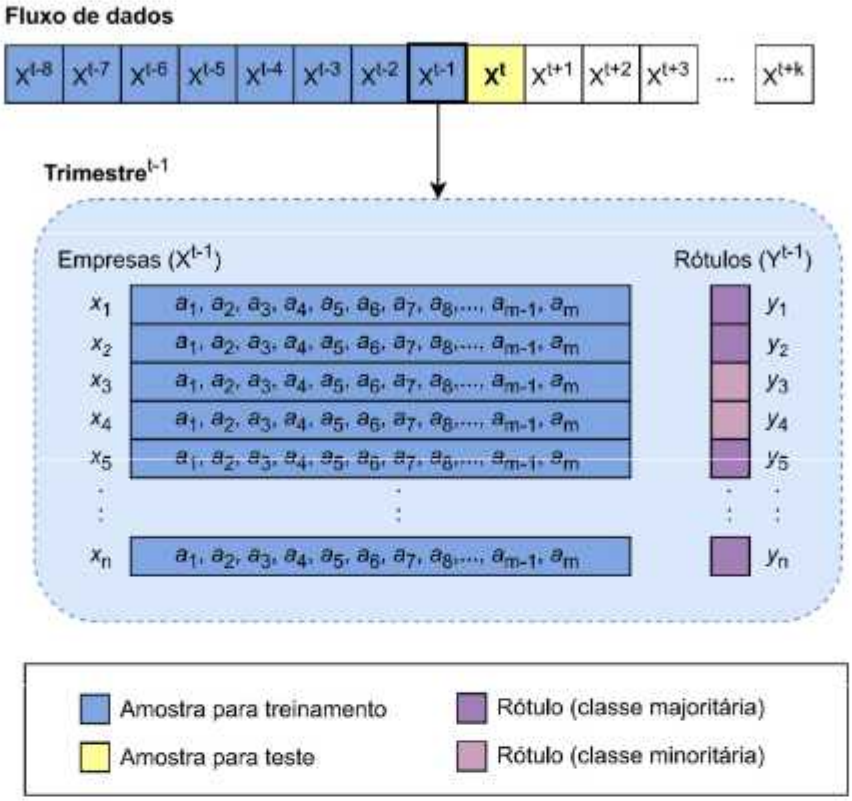


Figura 4.1: Infografo do fluxo de dados por trimestre.

Indicadores calculados a partir dessas contas podem apresentar problemas, por exemplo, *Giro de estoque* (ver Seção A.1.2 do apêndice) é o resultado da divisão do valor de *Custo de mercadorias vendidas* pelo valor do *Estoque*. Entretanto, em alguns casos o *Estoque* é zero, gerando uma divisão por zero, o que tende a gerar variáveis com valores inválidos. Portanto, indicadores nessa situação receberam o valor zero.

4.2 Processamento de dados não estacionários

Neste estudo, os dados são tratados como uma sequência de informações trimestrais $t-h, ..., t-2, t-1, t, t+1, t+2, ..., t+k$, onde $t-h$ é um momento passado (h número de trimestres passados), $t = 0$ equivale ao momento presente (atual) e $t+k$ é algum momento no futuro (k trimestres a receber). Em cada trimestre há um conjunto de dados de empresas distintas, onde cada registro tem 84 atributos, como ilustrado na Figura 4.1. É possível observar uma amostra de elementos para treinamento (em azul) e a amostra de teste (amarelo). Em um trimestre específico da amostra de treinamento, vê-se: os atributos de cada empresa (n empresas), os rótulos das classes majoritárias (roxo escuro) e rótulos das classes minoritárias (roxo claro), totalizando n rótulos. Lembrando, que a amostra de teste não tem rótulo, pois eles são preditos pelo modelo.

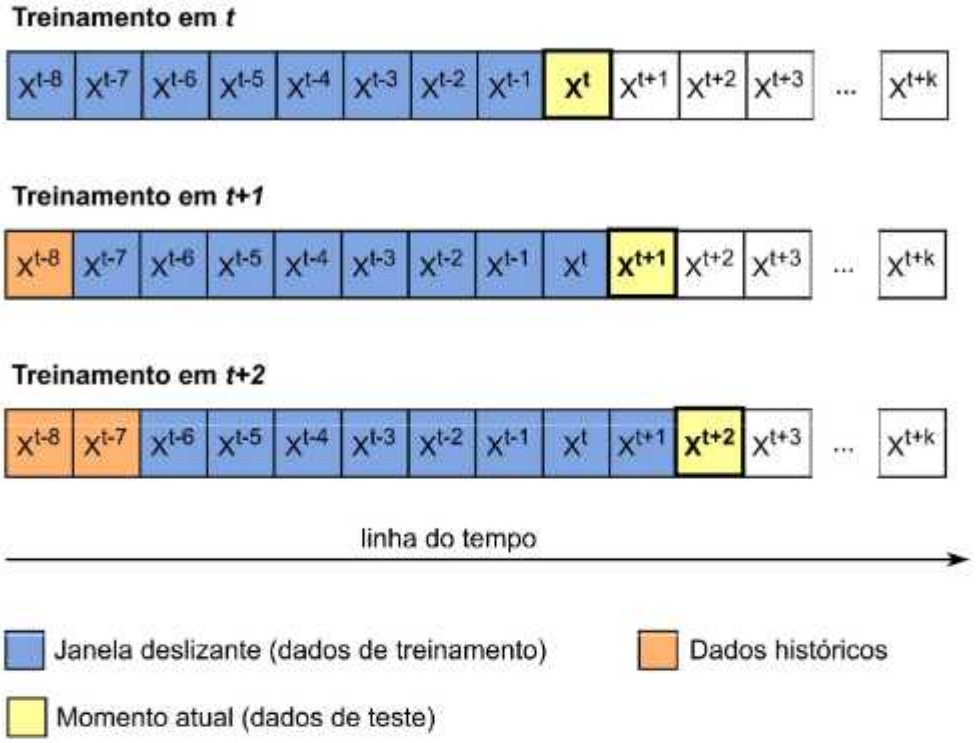


Figura 4.2: Funcionamento da janela deslizante e do histórico em 3 momentos distintos e consecutivos (t, $t+1$ e $t+2$).

Diante da dificuldade de mudança na distribuição dos dados no decorrer do tempo [26, 18, 90, 140], é utilizado o mecanismo de janela deslizante [44] para treinar o modelo com amostras de dados mais recentes (próximas de $t = 0$). Esse mecanismo busca minimizar o impacto de desvios de conceito no modelo.

4.2.1 Janela deslizante

Nos experimentos conduzidos por este estudo, a janela deslizante adotada tem 8 trimestres de tamanho (2 anos) e é testada com os dados do trimestre corrente ($t = 0$). O primeiro treinamento do modelo deve ocorrer quando a janela estiver cheia (dados dos 8 trimestres), a partir desse ponto, o treinamento deve ser refeito após remover o trimestre mais antigo ($t - 8$) e adicionar o trimestre mais recente ($t - 1$). Quando um trimestre é removido da janela deslizante ele é armazenado em uma base histórica. Entretanto, para diminuir o consumo de memória, apenas a informação da classe minoritária é armazenada (Figura 4.2). O histórico é utilizado para minimizar o problema de desbalanceamento, pois juntamente com as instâncias da classe minoritátia da janela deslizante, são utilizados para equilibrar as classes. O uso das instâncias da classe minoritária do histórico e da janela deslizante diminuem a geração de instâncias sintéticas.

4.2.2 Desbalanceamento

Como observado, no Capítulo 2, técnicas de balanceamento por meio de reamostragem como SMOTE e suas variantes tem sido a preferência nos últimos anos. Dessa forma, os desempenhos dos modelos serão comparado entre as seguintes técnicas: SMOTE [143], B-SMOTE [145], *Adaptive Synthetic* (ADASYN) [151], *Support Vector Machine-SMOTE* (SVM-SMOTE) [152], *SMOTE with Edited Nearest Neighbor* (SMOTE-ENN) [153] e *SMOTE with Tomek Links* (SMOTE-Tomek) [154].

A aplicação de variantes de SMOTE justifica-se porque a utilização de elementos sintéticos pode acarretar em problemas no treinamento do modelo [143]. Para minimizar esses problemas, as variantes tentam direcionar a geração dos elementos sintéticos de maneira que possam ser mais significativas ao invés de aleatória. Observa-se que as técnicas SMOTE-ENN e SMOTE-Tomek utilizam, de forma conjunta, a sobreamostragem e a subamostragem. Quando elas são aplicadas em conjunto, na medida correta, podem apresentar resultados melhores que a aplicação isolada das técnicas [143].

O histórico também é utilizado para reduzir a necessidade de criação de elementos sintéticos. Devido a utilização de elementos antigos e para evitar que eles influenciem negativamente o treinamento por conta de desvios de conceito, é adotado um mecanismo de esquecimento [142], como descrito a seguir.

$$f(t_h) = \begin{cases} 1 & \text{, se } t_h < 0 \\ e^{-\alpha \cdot t_h} & \text{, se } 0 \leq t_h < H \end{cases}$$

Onde, t_h é um trimestre do histórico e $t_h = 0$ é o trimestre mais recente e h é a distância do trimestre do histórico para janela deslizante. E α é o coeficiente de esquecimento, indicando a velocidade com que o algoritmo vai esquecer os dados, $0 \leq \alpha \leq 1$. Quando $\alpha = 0$ indica que não há esquecimento. Na Figura 4.3 é possível observar como o coeficiente impacta a curva de esquecimento. Assim, na Figura 4.3, quando a distância entre a janela deslizante é o trimestre é um ($h = 1$), apenas 36,8% das instâncias em DF são lembradas, isto é, 63,2% (1 - 0,368) são esquecidas.

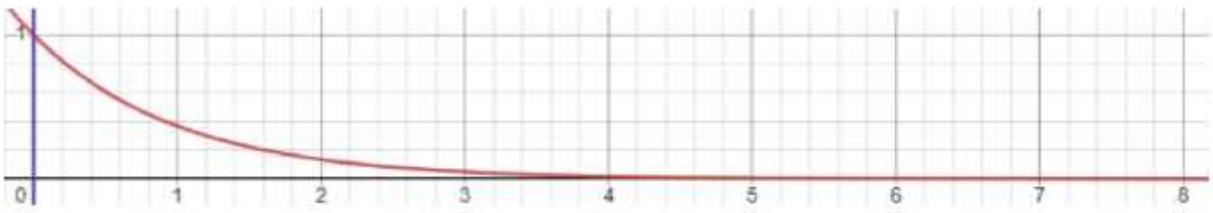


Figura 4.3: Curva de esquecimento com $\alpha = 1$, quando $t = 1$, $f(t) = 0,368$ (36,8%).

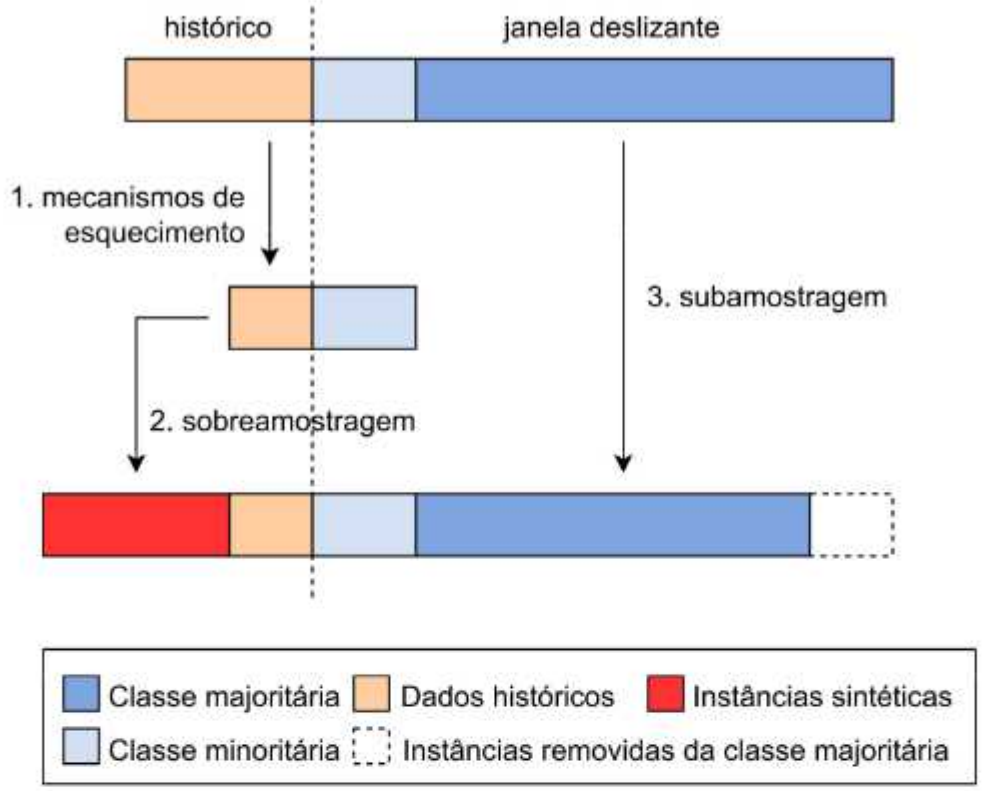


Figura 4.4: Processo de reequilíbrio da classes.

Na Figura 4.4 é possível visualizar o processo de reequilíbrio da classes. No passo 1, são selecionadas as instâncias da classe minoritária da janela deslizante. No passo 2, são selecionadas as instâncias do histórico usado o mecanismo de esquecimento. Então, é aplicada a sobreamostragem (*i.e.* B-SMOTE) no conjunto de instâncias formado pelos elementos selecionados nos passos 1 e 2. Por fim, no passo 3 e antes do treinamento, é aplicado o método de subamostragem para reduzir o conjunto majoritário, visando reduzir o desbalanceamento entre as classes [143].

Para ilustrar a utilização desses mecanismos, por simplificação, assume-se uma janela deslizante de 8 trimestres, um histórico de 8 trimestres e duas classes de elementos, a classe M (majoritária) e a classe m (minoritária). Sabendo que em cada trimestre existe um desbalanceamento 1:10 e, portanto, que todos os trimestres tem 1000 instâncias da classe M e 100 instâncias da classe m. Toda a janela terá instâncias da classe M e instâncias da classe m, totalizando 1100 instâncias. Para diminuir esse desequilíbrio utiliza-se instâncias da classe m armazenadas no histórico, observando a taxa de esquecimento com $\alpha = 1$, como descrito na Figura 4.3. Cada trimestre do histórico tem apenas 100 instâncias da classe m e não tem instâncias da classe M. Assim, quando o histórico estiver completo com os 8 trimestres, ele terá um total de 800 instâncias da classe m, entretanto, devido ao mecanismo de esquecimento apenas 158 instâncias serão utilizadas (passo 1 da Figura 4.4).

As instâncias selecionadas do histórico serão concatenadas com as instâncias minoritárias da janela deslizante resultando em um conjuntos de 958 instâncias (158 + 800), que no passo 2 (Figura 4.4) serão usadas na criação de instâncias sintéticas para equilibrar as classes.

4.3 Avaliação

Mudando o modelo de treinamento é possível comparar os principais algoritmos de AM identificados durante a etapa de revisão sistemática, entre eles: RL, SVM [155, 156],RF [157], AD[158], Extreme Gradient Boosting (XGBoost) [159, 160] e Categorial Boosting (CatBoost) [161]. A Tabela 4.2 apresenta os hiperparâmetros utilizados na inicialização dos modelos preditivos, na maioria dos casos foram utilizados os valores padrões dos parâmetros de cada classificador. Com exceção de RL que especificou `solver='liblinear'` e `max_iter=300`, e SVM que utilizou `probability=True`.

Os registros das empresas são rotulados em DF ou SDF de acordo com as informações da CVM no trimestre em que foi observada a situação. A verificação do modelo é feita por meio do horizonte preditivo (k), que deve variar entre os valores 2, 4, 8, 12, 16, 20 e 24. Por exemplo, $k = 4$ significa que o modelo será validado com o trimestre $t + 4$, assim a identificação de falência estará ocorrento 1 ano antes da mesma ter sido informada na CVM.

Quando se trabalha dados sequenciais, a validação cruzada não é trivial. Não é possível escolher amostras aleatórias e atribuí-las ao conjunto de avalidação ou ao conjunto de treinamento porque não faz sentido usar os valores do futuro para prever valores no passado. Em outras palavras, é necessário evitar olhar para o futuro durante o treinamento do modelo, pois existe uma dependência temporal entre as observações e deve-se preservar essa relação durante a avaliação (teste).

Nesses casos, é utilizado um método de validação contínua, como a validação cruzada aninhada para séries temporais (*nested cross-validation on time series*) [162], ilustrada na Figura 4.5. Primeiramente ($t = 0$), gera-se um conjunto de treinamento a partir de dados da janela deslizante, do histórico (após mecanismos de esquecimento) e das instâncias sintéticas. Separam-se os trimestres relativos ao horizonte preditivo e, o trimestre seguinte, é selecionado como conjunto avaliação (teste). Na etapa seguinte ($t + 1$), ocorre um deslocamento de um trimestre do conjunto de teste para o horizonte preditivo, à janela deslizante e ao histórico (com geração de novas instâncias sintéticas), resultando em um novo conjunto de treinamento e teste.

Na Figura 4.5, observa-se que, o conjunto de treinamento é composto por dados da janela deslizante em azul escuro (classe majoritária) e azul claro (classe minoritária),

Tabela 4.2: Hiperparâmetros utilizados para treinamento.

Classificador	**Hiperparâmetros**
RL	`penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, solver='liblinear', max_iter=300, multi_class='auto', verbose=0, warm_start=False`
AD	`criterion='gini', splitter='best', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, min_impurity_decrease=0.0, ccp_alpha=0.0`
SVM	`C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=True, tol=0.001, cache_size=200, verbose=False, max_iter=-1, decision_function_shape='ovr', break_ties=False`
RF	`n_estimators=100, criterion='gini', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='sqrt', min_impurity_decrease=0.0, bootstrap=True, oob_score=False, verbose=0, warm_start=False, ccp_alpha=0.0`
XGBoost	`eta=0.3, gamma=0, max_depth=6, min_child_weight=1, max_delta_step=0, subsample=1, sampling_method='uniform', colsample_bytree=1, colsample_bylevel=1, colsample_bylevel=1, lambda=1, alpha=0, tree_method='auto', scale_pos_weight=1, refresh_leaf=1, process_type='default', grow_policy='depthwise', max_leaves=0, max_bin=256, num_parallel_tree=1, multi_strategy='one_output_per_tree'`
CatBoost	All parameters set to `None`.

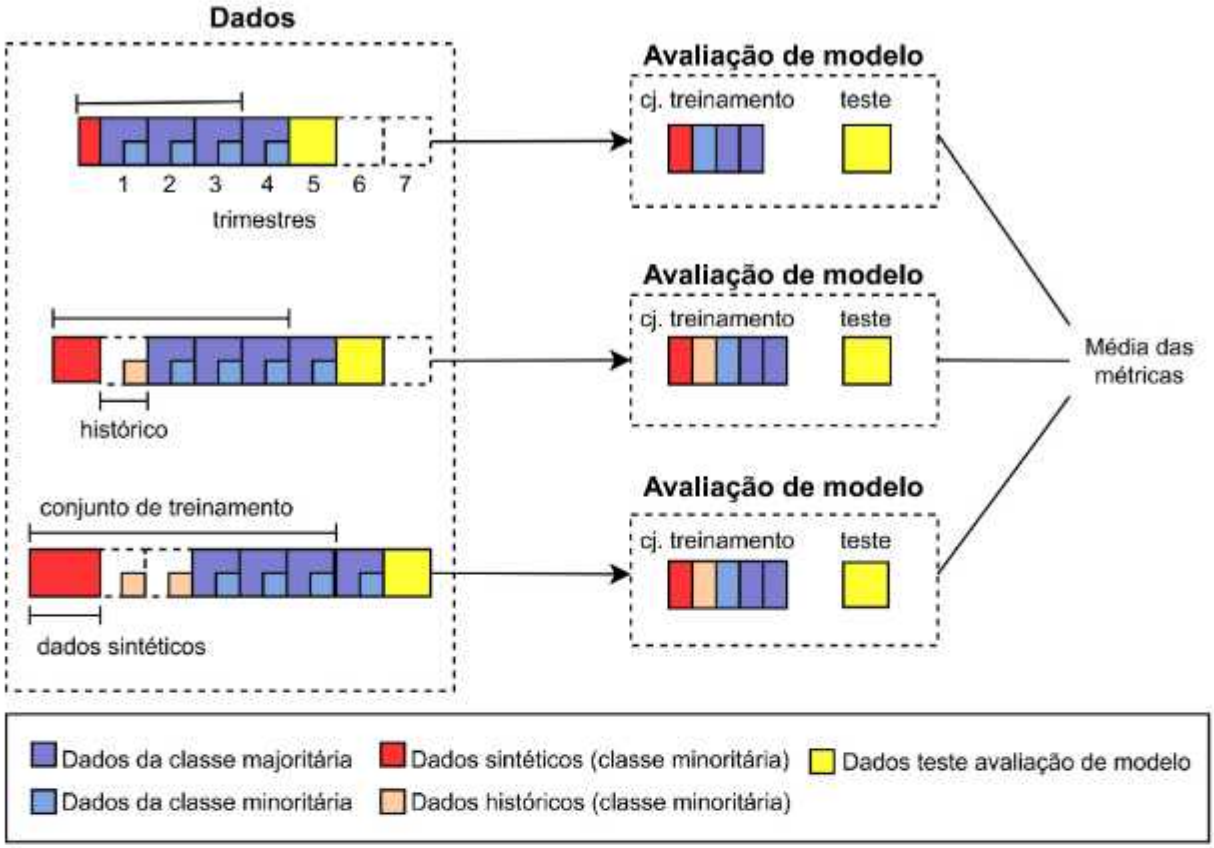


Figura 4.5: Aninhamento de métodos de validação cruzada para séries temporais.

dados históricos da classe minoritária (em salmão) e dados sintéticos (em vermelho). Eles serão utilizados para treinar o modelo, que será avalidado através de um conjunto de testes (em amarelo). Esse processo deve ser repetido em diferentes momentos (trimestres), sempre utilizando o trimestre corrente como conjunto de teste. Ao atingir uma quantidade satisfatória de repetições é calculada a média das métricas e elaborada uma curva de evolução de cada uma.

Na implementação dos experimentos descritos neste estudo foram utilizada bibliotecas já consolidadas, como: *scikit-learn*[6], XGBoost[7] e CatBoost[8]. Assim, como a biblioteca *imbalanced-learn*[9] para tratamento do desbalanceamento de dados, com implementação do SMOTE, demais variantes (B-SMOTE, ADASYN, SVM-SMOTE, SMOTE-ENN e SMOTE-Tomek) e métricas de avaliação utilizadas neste estudo (Precisão, Sensibiliade, F_1-Score, G_{mean}, AUC-ROC e AUC-PS). Os modelos preditivos citados foram aplicados utilizando hiperparâmetros padrões de cada método fornecido pela API, com exceção da RL que utilizou `solver='liblinear'` e `max_iter=300` e SVM que utilizou

[6] `https://scikit-learn.org/`
[7] `https://xgboost.readthedocs.io/`
[8] `https://catboost.ai/`
[9] `https://imbalanced-learn.org/`

probability=True.

Capítulo 5

Resultados

Os resultados obtidos devem servir como evidências para validar as hipóteses de pesquisa discutidas na Seção 1.3. Eles permitirão identificar a melhor forma de tratar a questão do desbalanceamento de dados e identificar os algoritmos capazes de predizer a DF com maior antecedência e com melhor desempenho, no contexto deste estudo. Os experimentos foram realizados por meio da variação dos algoritmos de classificação (RL, AD, SVM, RF, XGBoost e CatBoost), algoritmos de balanceamento (SMOTE, B-SMOTE, ADASYN, SVM-SMOTE, SMOTE-ENN e SMOTE-Tomek), taxa de desbalanceamento (0%, 50% e 100%) e do horizonte preditivo (2, 4, 8, 12, 16, 20 e 24) em trimestres.

Este capítulo está dividido em duas partes. A Seção 5.1 descreve a base de dados de indicadores econômico-financeiros que foi produzida a partir de dados da CVM. A Seção 5.2 apresenta os experimentos de predição de DF, sua organização e os resultados.

5.1 Base de dados

Este estudo produziu uma base de dados de indicadores econômico-financeiros com marca temporal trimestral. Essa base de dados basea-se em informações fornecidas a CVM pelas empresas listadas na bolsa de valores brasileira (B3 [1]). Então, a partir dos arquivos contáveis disponíveis no Portal de Dados Abertos da CVM foi possível extraír um total de 84 indicadores, que foram marcados com a informação de trimestres na coluna `QUARTER`, podendo assumir os valores `[ANO]-03-31`, `[ANO]-06-30`, `[ANO]-09-30` e `[ANO]-12-31` para o 1º, 2º, 3º e 4º trimestres, respectivamente, onde `[ANO]` é a informação do ano do trimestre, neste conjunto de dados ele pode variar de 2011 a 2020. A informação de CNPJ das empresas foi substituída por um valor aleatório, a fim de não permitir a identificação das partes, na coluna `ID`. A informação de rótulo de dada resgistro está na coluna `LABEL`, assumindo os valores 0 quando a empresa não está em DF e 1 quando a

[1] `https://b3.com.br/`

empresa encontra-se em DF. Os outros atributos (Atr.) ou indicadores foram nomeados com `A[NUMERO]`, onde o `[NUMERO]` varia de 1 a 84, como descrito na Tabela 5.1. A forma de calculo de cada um desses atributos está descrita no Apêndice A.1.

Ao final do processo de extração de dados foi obtida uma base de dados com 23.834 registros, sendo 23.183 registros de empresas SDF e 651 registros de empresas em DF, representado em percentuais como 97,27% e 2,73%, respectivamente. Complementarmente, a Figura 5.1 apresenta um gráfico de colunas empilhadas, com o a quantidade de empresas no eixo vertical e cada trimestre do período de 2011 a 2020 no eixo horizontal. A parte vermelha das colunas é a quantidade de empresas SDF e a parte azul são as empresa em DF. Portanto, o desbalanceamento está distribuído em cada trimestre da base de dados a uma taxa elevada, caracterizando o conjunto como fortemente desbalanceado.

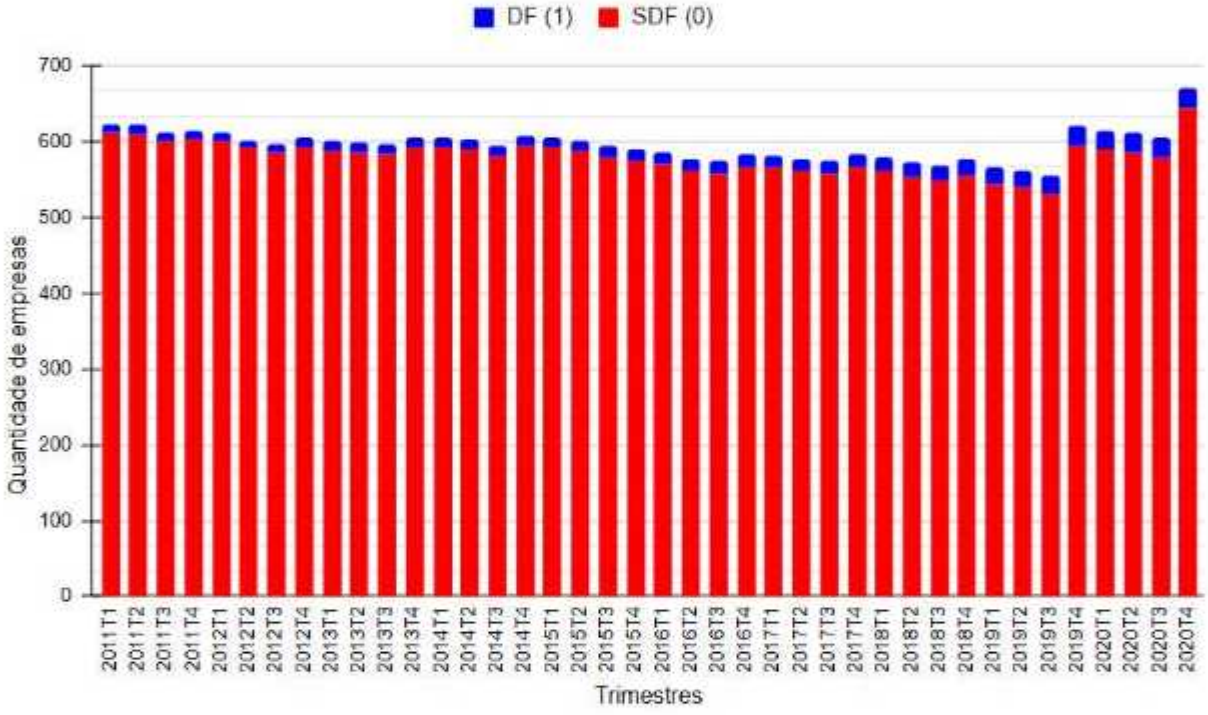


Figura 5.1: Infografo de empresa em DF e empresas em situação normal.

Por fim, a base de dados está disponível no Github[2] para acesso público. Ela também foi utilizada para realizar experimentos a fim de comparar os modelos e verificar a viabilidade de uso da base de dados.

[2] `https://github.com/rubensmchaves/ml-fdp`

Tabela 5.1: Atributos da base de dados de indicadores econômico-financeiros.

Atr.	Descrição	Atr.	Descrição
A1	Ativo total	A43	Relação de obrigações e ativos tangíveis
A2	Ativo circulante	A44	Índice de liquidez
A3	Disponibilidades	A45	Índice de ativos recebíveis
A4	Recebíveis	A46	Índice de ativos fixos
A5	Estoque	A47	Índice de A17 para A9
A6	Ativos de longo prazo	A48	Índice de dívida atual
A7	Ativos intangíveis	A49	Margem de lucro
A8	Ativos tangíveis	A50	Relação recebíveis e receita
A9	Ativos fixos	A51	Relação estoque e lucro
A10	Depreciação acumulada	A52	Rotação de estoque
A11	Amortização acumulada	A53	Rotação de contas a pagar
A12	Investimentos	A54	Rotação de ativo circulante
A13	Passivo total	A55	Relação ativo permanente e lucro
A14	Passivo circulante	A56	Rotação do capital
A15	Passivo não-circulante	A57	Retorno sobre ativos
A16	Exigíveis (A13 - A14)	A58	Relação lucro líquido e ativos totais
A17	Patrimônio líquido (A12 - A15)	A59	Relação A31 e A2
A18	Capital social	A60	Relação lucro líquido e ativos permanente
A19	Reservas	A61	Retorno sobre patrimônio
A20	Provisões	A62	Margem operacional
A21	Empréstimos de longo prazo	A63	Relação custo operacional total e A22
A22	Receita bruta	A64	Relação de despesas e receitas
A23	Despesas	A65	Índice de despesas de gestão
A24	Lucro líquido	A66	Índice de endividamento
A25	Despesas operacionais	A67	Fluxo de Caixa Livre
A26	Lucro operacional	A68	Relação A32 e A24
A27	Resultado financeiro	A69	Relação de A32 e A22
A28	Despesas financeiras	A70	Taxa de recuperação de caixa
A29	Lucro antes de impostos	A71	Grau de alavancagem financeira
A30	Despesas com impostos	A72	Grau de alavancagem operacional
A31	Resultado líquido	A73	Grau de alavancagem combinada
A32	Fluxo de caixa operacional	A74	Taxa de cresc. de manutenção de capital
A33	Fluxo de caixa de investimentos	A75	Taxa de cresc. de capital acumulado
A34	Fluxo de caixa de financiamento	A76	Taxa de cresc. de ativos totais
A35	Ações em circulação	A77	Taxa de cresc. de A61
A36	Liquidez corrente	A78	Taxa de cresc. de lucro líquido
A37	Liquidez seca	A79	Taxa de cresc. de lucro operacional
A38	Liquidez imediata	A80	Taxa de cresc. de receitas operacionais
A39	Índice de cobertura de juros	A81	Taxa de cresc. de custos operacionais
A40	Índice de endividamento total	A82	Lucro por ação
A41	Índice de cobertura de ativos tangíveis	A83	Valor de ativos líquidos por ação
A42	Relação patrimônio e dívida	A84	Caixa livre por ação
ID: Identificador aleatório.			
QUARTER: Marca temporal (trimestral).			
LABEL: 0 - SDF; 1 - DF			

5.2 Experimentos

Os resultados foram obtidos a partir de experimentos combinados entre classificadores (RL, AD, SVM, RF, XGBoost e CatBoost), técnicas de reamostragem (SMOTE, B-SMOTE, ADASYN, SVM-SMOTE, SMOTE-ENN e SMOTE-Tomek), taxas (0%, 50% e 100%) e horizonte preditivo (2, 4, 8, 12, 16, 20 e 24) em trimestres. A combinação de classificadores (6) e horizontes preditivos (7) resultou em 42 experimentos. Com o uso das técnicas de reamostagem, a combinação de classificadores (6), técnicas de reamostragem (6), taxas de balanceamento (2) e horizontes preditivos (7) resultou em 504 experimentos. Somando-se esse resultados, ao todo foram realizados 546 experimentos. Para minimizar os efeitos de desvio de conceito foi utilizado a técnica de janela deslizante com o tamanho de 8 trimestres (2 anos) e um mecanismo de esquecimento com coeficiente de esquecimento igual a 1. Os resultados foram gerados utilizando o método de validação cruzada aninhada em séries temporais, descrito no Capítulo 4.

A partir dos resultados da combinação inicial (classificadores, técnica de reamostragem e taxa de balanceamento), com cada horizonte preditivo, calculou-se a média entre todos os horizontes para cada métrica (Precisão, Sensibilidade, F_1-Score, G_{mean}, AUC-ROC e AUC-PS). Essas médias estão organizada na Tabela 5.2, onde a primeira coluna contém a informação da métrica de avaliação (Métricas), a segunda e a terceira combinam as informações de técnica de reamostragem (Reamostragem) e a taxa de balanceamento utilizada (Tx), enquanto as demais organizam os resultados pelos classificadores. Quando os resultados de todos os modelos são avaliados em conjunto para uma determinada métrica, o melhor resultado entre eles é destacado em negrito. Por sua vez, quando a avaliação de uma métrica ocorre para o subconjunto de resultados de um modelo específico o melhor resultado é destacado em itálico. Por exemplo, na métrica Precisão, o melhor resultado geral é 0,9422, apresentado pelo RF, destacado em negrito. Quando analisamos o melhor resultado de Precisão para o CatBoost o melhor resultado é 0,9024, destacado em itálico.

Observando os resultados por classificador em relação às métricas (em negrito), é possível perceber que o classicador CatBoost, das seis métricas utilizadas, conseguiu o melhor resultado em 3 delas (3/6), F_1-Score, AUC-ROC e AUC-PS. Enquanto que, os classificadores SVM, RF e XGBoost conseguiram o melhor resultado em apenas 1 de 6 métricas (1/6), Sensibilidade, Precisão e G_{mean}, respectivamente. Os classificadores de RL e AD não foram melhor em nenhuma delas (0/6).

Os classificadores estão listados na Tabela 5.3, ordenados do melhor para o pior. A primeira coluna (#) apresenta a ordem de classificação, a segunda o nome do classificador, a terceira as métricas em que o classificador foi melhor e a quarta (Destaques) totaliza

Tabela 5.2: Resultados de classificadores utilizando técnicas de balanceamento (taxas de 0, 0,5 and 1).

Métrica	Balanceamento	Tx	RL	AD	SVM	RF	XGBoost	CatBoost
Precisão	-	0,00	0,0794±0,01	0,4110±0,11	0,0000±0,00	**0,9422±0,03**	*0,8166±0,09*	*0,9021±0,07*
	SMOTE	0,5	0,1466±0,01	0,3890±0,07	0,0048±0,01	0,8175±0,06	0,6157±0,08	0,8036±0,07
		1,0	0,1767±0,01	0,3800±0,07	0,0367±0,00	0,7806±0,05	0,5978±0,06	0,7620±0,07
	B-SMOTE	0,5	0,1413±0,01	0,3565±0,10	0,0069±0,01	0,8790±0,06	0,6332±0,12	0,7762±0,08
		1,0	0,1608±0,01	0,3670±0,10	0,0343±0,00	0,8394±0,08	0,6102±0,10	0,7406±0,09
	ADASYN	0,5	0,1207±0,01	0,3546±0,12	0,0138±0,01	0,8097±0,09	0,6102±0,10	0,7897±0,08
		1,0	0,1468±0,01	0,3608±0,12	*0,0472±0,03*	0,7716±0,12	0,5817±0,09	0,7363±0,10
	SVM-SMOTE	0,5	0,1339±0,02	0,4051±0,06	0,0073±0,01	0,8157±0,06	0,6243±0,07	0,7976±0,07
		1,0	0,1688±0,01	*0,4216±0,08*	0,0367±0,00	0,7776±0,06	0,5792±0,06	0,7407±0,09
	SMOTE-ENN	0,5	0,1453±0,02	0,3924±0,09	0,0055±0,01	0,8155±0,05	0,6237±0,09	0,8039±0,06
		1,0	0,1801±0,01	0,3863±0,07	0,0349±0,00	0,7903±0,05	0,5956±0,06	0,7640±0,07
	SMOTE-Tomek	0,5	0,1378±0,02	0,4034±0,09	0,0048±0,01	0,8298±0,06	0,6158±0,08	0,7908±0,07
		1,0	*0,1805±0,01*	0,3715±0,08	0,0367±0,00	0,7839±0,05	0,5867±0,06	0,7507±0,07
Sensibilidade	-	0,00	0,0867±0,00	0,3316±0,13	0,0000±0,00	0,2757±0,12	0,2983±0,15	0,3236±0,16
	SMOTE	0,5	0,3481±0,08	0,3367±0,12	0,0007±0,00	0,2811±0,15	0,4580±0,17	0,4562±0,17
		1,0	0,5256±0,07	0,2959±0,14	**0,9768±0,01**	0,2575±0,15	0,4897±0,17	0,4707±0,18
	B-SMOTE	0,5	0,3369±0,06	0,3100±0,13	0,0011±0,00	0,2509±0,15	0,3834±0,16	0,4100±0,19
		1,0	0,4703±0,05	0,3035±0,13	0,8853±0,03	0,2359±0,15	0,3941±0,16	0,4124±0,19
	ADASYN	0,5	0,3408±0,05	0,3072±0,15	0,0022±0,00	0,2551±0,18	0,4043±0,18	0,4152±0,20
		1,0	0,4957±0,07	0,2970±0,14	0,6652±0,16	0,2324±0,17	0,4238±0,18	0,4199±0,21
	SVM-SMOTE	0,5	0,3439±0,09	*0,3491±0,13*	0,0011±0,00	*0,3292±0,17*	0,4776±0,16	0,4658±0,17
		1,0	0,5320±0,08	0,3404±0,15	0,9760±0,01	0,3171±0,17	*0,5073±0,18*	*0,4917±0,18*
	SMOTE-ENN	0,5	0,3326±0,09	0,3370±0,12	0,0009±0,00	0,2929±0,15	0,4581±0,16	0,4633±0,18
		1,0	0,5275±0,06	0,3137±0,13	0,9157±0,04	0,2643±0,15	0,4956±0,17	0,4888±0,17
	SMOTE-Tomek	0,5	0,3289±0,09	0,3348±0,12	0,0007±0,00	0,2869±0,16	0,4583±0,17	0,4518±0,17
		1,0	*0,5399±0,05*	0,2909±0,13	0,9768±0,01	0,2617±0,15	0,4889±0,17	0,4736±0,17
F_1-Score	-	-	0,0812±0,00	0,3562±0,12	0,0000±0,00	0,4133±0,03	0,4203±0,14	0,4581±0,14
	SMOTE	0,5	0,2019±0,03	0,3516±0,10	0,0013±0,00	0,3976±0,03	0,5132±0,12	0,5683±0,12
		1,0	0,2611±0,02	0,3138±0,12	0,0705±0,00	0,3640±0,03	0,5247±0,12	0,5665±0,13
	B-SMOTE	0,5	0,1912±0,02	0,3200±0,12	0,0019±0,00	0,3658±0,03	0,4649±0,13	0,5179±0,15
		1,0	0,2319±0,01	0,3231±0,12	0,0656±0,00	0,3442±0,03	0,4657±0,13	0,5124±0,15
	ADASYN	0,5	0,1730±0,01	0,3187±0,14	0,0038±0,00	0,3605±0,03	0,4729±0,14	0,5241±0,15
		1,0	0,2211±0,01	0,3143±0,13	0,0558±0,01	0,3301±0,04	0,4769±0,13	0,5165±0,16
	SVM-SMOTE	0,5	0,1885±0,03	*0,3614±0,10*	0,0020±0,00	*0,4474±0,02*	*0,5902±0,12*	0,5743±0,13
		1,0	0,2524±0,02	0,3602±0,13	*0,0706±0,00*	0,4273±0,03	0,5273±0,12	0,5792±0,13
	SMOTE-ENN	0,5	0,1977±0,03	0,3517±0,11	0,0016±0,00	0,4108±0,03	0,5167±0,12	0,5719±0,13
		1,0	0,2649±0,01	0,3307±0,11	0,0671±0,00	0,3730±0,03	0,5288±0,12	**0,5812±0,12**
	SMOTE-Tomek	0,5	0,1896±0,03	0,3521±0,11	0,0013±0,00	0,4030±0,03	0,5142±0,12	0,5621±0,12
		1,0	*0,2674±0,01*	0,3113±0,11	0,0705±0,00	0,3700±0,03	0,5216±0,12	0,5681±0,12
G_{mean}	-	-	0,2490±0,09	0,3575±0,11	0,0000±0,00	0,5124±0,11	0,5272±0,14	0,5518±0,14
	SMOTE	0,5	0,5429±0,06	0,3604±0,10	0,0038±0,01	0,5089±0,14	0,6614±0,12	0,6624±0,12
		1,0	0,6837±0,12	0,5190±0,13	*0,2334±0,06*	0,4838±0,14	0,6845±0,12	0,6722±0,13
	B-SMOTE	0,5	0,5226±0,04	0,5346±0,11	0,0057±0,01	0,4744±0,15	0,6008±0,13	0,6218±0,15
		1,0	0,6320±0,13	0,5298±0,12	0,2055±0,07	0,4572±0,15	0,6103±0,13	0,6236±0,15
	ADASYN	0,5	0,5147±0,03	0,5266±0,13	0,0111±0,01	0,4714±0,18	0,6171±0,14	0,6248±0,15
		1,0	0,6474±0,13	0,5193±0,13	0,1552±0,05	0,4448±0,18	0,6325±0,13	0,6275±0,16
	SVM-SMOTE	0,5	0,5238±0,06	*0,5698±0,11*	0,0057±0,01	*0,5544±0,14*	0,6766±0,12	0,6698±0,13
		1,0	0,6843±0,12	0,5620±0,12	0,2316±0,06	0,5407±0,15	**0,6961±0,12**	*0,6883±0,13*
	SMOTE-ENN	0,5	0,5307±0,06	0,5596±0,11	0,0047±0,01	0,5218±0,14	0,6620±0,12	0,6668±0,13
		1,0	0,6842±0,12	0,5379±0,12	0,2284±0,06	0,4907±0,14	0,6908±0,12	0,6865±0,12
	SMOTE-Tomek	0,5	0,5228±0,06	0,5575±0,10	0,0038±0,01	0,5137±0,14	0,6618±0,12	0,6593±0,12
		1,0	*0,6927±0,11*	0,5161±0,12	0,2321±0,06	0,4871±0,14	0,6834±0,12	0,6754±0,12
AUC-ROC	-	-	0,7525±0,01	0,6571±0,07	0,5106±0,03	0,9102±0,03	0,9405±0,02	0,9436±0,03
	SMOTE	0,5	0,8004±0,02	0,6590±0,06	0,6035±0,07	*0,9298±0,03*	0,9379±0,03	0,9484±0,02
		1,0	0,8304±0,03	0,6397±0,07	0,6461±0,01	0,9280±0,03	0,9395±0,03	0,9514±0,02
	B-SMOTE	0,5	0,7832±0,02	0,6448±0,06	0,6027±0,03	0,9223±0,03	0,9331±0,03	0,9471±0,03
		1,0	0,8073±0,02	0,6424±0,07	0,5932±0,03	0,9223±0,03	0,9304±0,03	0,9469±0,03
	ADASYN	0,5	0,7859±0,02	0,6440±0,08	0,6279±0,05	0,9205±0,03	0,9315±0,03	0,9469±0,03
		1,0	0,8145±0,02	0,6395±0,07	0,6051±0,02	0,9189±0,04	0,9310±0,03	0,9481±0,02
	SVM-SMOTE	0,5	0,7950±0,03	*0,6655±0,06*	0,5943±0,07	0,9291±0,02	0,9364±0,03	0,9474±0,03
		1,0	0,8268±0,03	0,6620±0,07	*0,6520±0,01*	0,9251±0,03	0,9374±0,03	0,9479±0,03
	SMOTE-ENN	0,5	0,7990±0,02	0,6595±0,06	0,5962±0,08	0,9269±0,03	0,9404±0,03	0,9506±0,02
		1,0	0,8306±0,02	0,6483±0,07	0,6410±0,01	0,9287±0,03	*0,9424±0,03*	**0,9520±0,02**
	SMOTE-Tomek	0,5	0,7959±0,03	0,6586±0,06	0,5986±0,07	0,9280±0,03	0,9384±0,03	0,9485±0,02
		1,0	*0,8314±0,02*	0,6370±0,07	0,6451±0,00	0,9276±0,03	0,9389±0,03	0,9513±0,02
AUC-PS	-	-	0,0768±0,00	0,3831±0,11	0,0349±0,00	0,5821±0,14	0,5677±0,14	0,6130±0,15
	SMOTE	0,5	0,1134±0,01	0,3745±0,09	0,0503±0,01	0,5733±0,13	0,5657±0,14	0,6352±0,14
		1,0	0,1363±0,01	0,3504±0,10	0,0646±0,00	0,5368±0,14	0,5768±0,13	0,6342±0,13
	B-SMOTE	0,5	0,1029±0,01	0,3454±0,11	0,0475±0,00	0,5725±0,14	0,5250±0,15	0,6067±0,15
		1,0	0,1195±0,01	0,3475±0,11	0,0617±0,02	0,5552±0,15	0,5131±0,15	0,5954±0,15
	ADASYN	0,5	0,0996±0,01	0,3432±0,13	0,0481±0,00	0,5522±0,16	0,5190±0,15	0,6039±0,16
		1,0	0,1175±0,01	0,3413±0,13	*0,0765±0,02*	0,5282±0,17	0,5083±0,14	0,5919±0,16
	SVM-SMOTE	0,5	0,1078±0,02	0,3886±0,09	0,0473±0,01	*0,5872±0,14*	0,5724±0,14	0,6388±0,14
		1,0	0,1311±0,01	*0,3926±0,11*	0,0764±0,01	0,5631±0,14	0,5816±0,14	0,6360±0,14
	SMOTE-ENN	0,5	0,1134±0,02	0,3764±0,10	0,0494±0,01	0,5789±0,13	0,5719±0,14	**0,6414±0,13**
		1,0	0,1383±0,01	0,3621±0,09	0,0680±0,01	0,5438±0,13	*0,5822±0,14*	0,6386±0,13
	SMOTE-Tomek	0,5	0,1096±0,02	0,3809±0,10	0,0498±0,01	0,5749±0,13	0,5691±0,14	0,6336±0,13
		1,0	*0,1586±0,01*	0,3437±0,10	0,0667±0,01	0,5438±0,13	0,5787±0,13	0,6335±0,13

a quantidade de métricas em que o classificador foi melhor. Então, percebe-se que o classificador CatBoost apresentou melhor desempenho.

Tabela 5.3: Ordenação de classificadores por quantidade de melhor resultados por métrica.

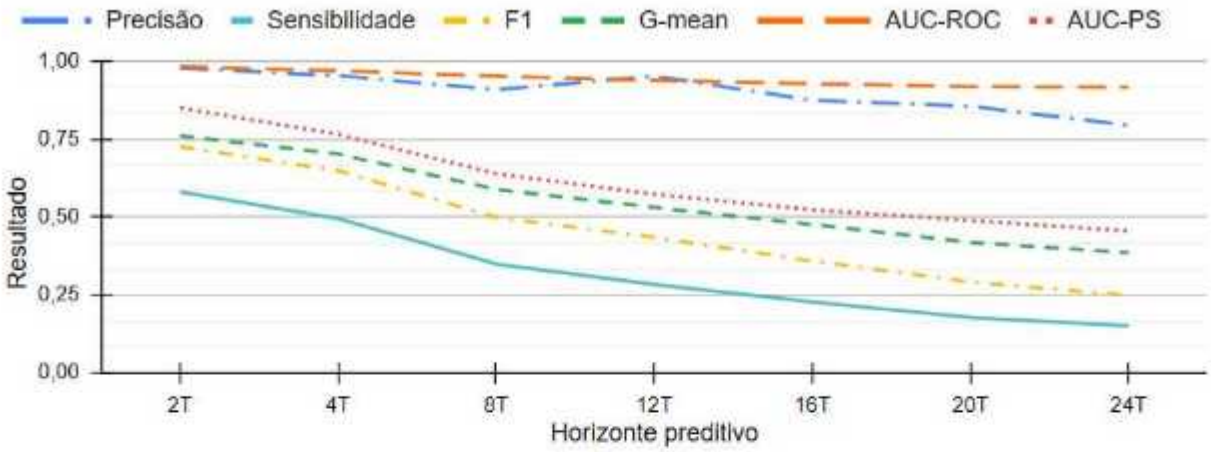
#	**Classificador**	**Métricas**	**Destaques**
1	CatBoost	F_1-Score AUC-ROC AUC-PS	3
2	SVM RF XGBoost	Sensibilidade Precisão G_{mean}	1
3	RL AD	-	0

De maneira análoga, observa-se pela Tabela 5.2 que os melhores resultados da coluna CatBoost (valores em itálico e/ou negrito), em cada métrica, identificam que a técnica de reamostragem SMOTE-ENN é mais adequada, pois consegue ser melhor em 3 das 6 métricas (F_1-Score, AUC-ROC e AUC-PS). Enquanto que, a técnica de reamostragem SVM-SMOTE consegue ser melhor em 2 de 6 (Sensibilidade e G_{mean}). Entretanto, para seleção da melhor configuração do *pipeline*, ainda é necessário responder a pergunta: qual taxa de balanceamento contribui para um melhor funcionamento do modelo?

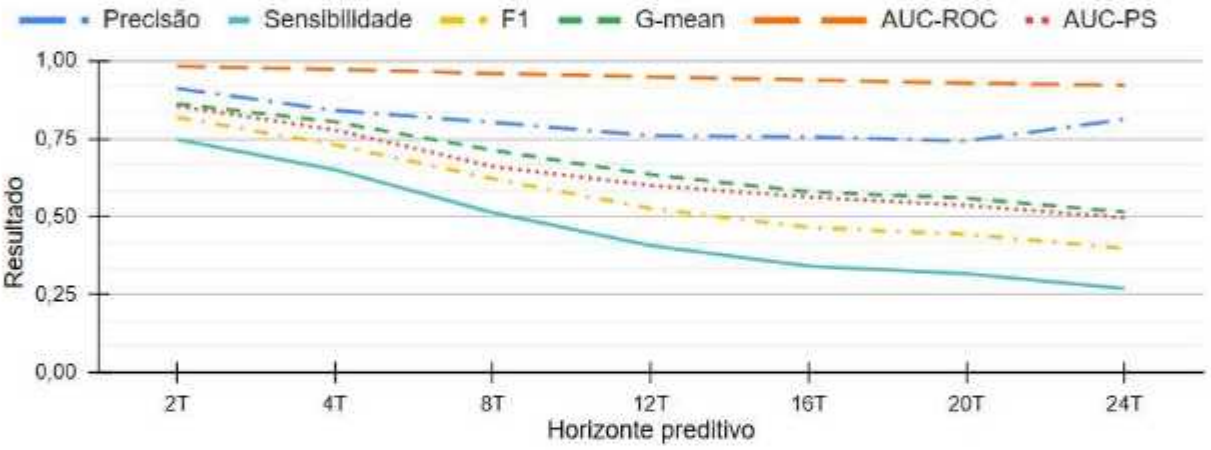
A resposta à pergunta vem da análise de gráficos de evolução do modelo à medida que o horizonte preditivo aumenta. Na Figura 5.2 são apresentados três gráficos de métricas que ilustram bem o comportamento do modelo CatBoost com SMOTE-ENN usando as taxas de 0% (Figura 5.2a), 50% (Figura 5.2b) e 100% (Figura 5.2c). Nessas figuras o eixo horizontal representa os valores de horizonte preditivo em trimestres (T) e o eixo vertical os resultados obtidos por cada métrica.

Os gráficos ilustrados na Figura 5.2 permitem observar que o aumento de horizonte predito tem um forte impacto no desempenho do modelo. Entretanto, a utilização da reamostragem mostrou-se capaz de minimizar o impacto do aumento do horizonte preditivo. É possível perceber que a inclinação negativa das curvas é atenuada. Por exemplo, a Sensibilidade, quando o horizonte é de 4 trimestres (4T), assume os valores de 0,4956, 0,6506 e 0,6658 para as taxas de reamostragem de 0%, 50% e 100%, respectivamente. Portanto, após o reequilíbrio das classes, a Sensibilidade passa a cruzar a linha de 0,5 após o horizonte preditivo de 8 trimestres (8T), permitindo identificar a DF com mais antecedência e mantendo o mesmo desempenho. Tendência semalhante também é observada nas métricas F_1-Score, G_{mean} e AUC-PS.

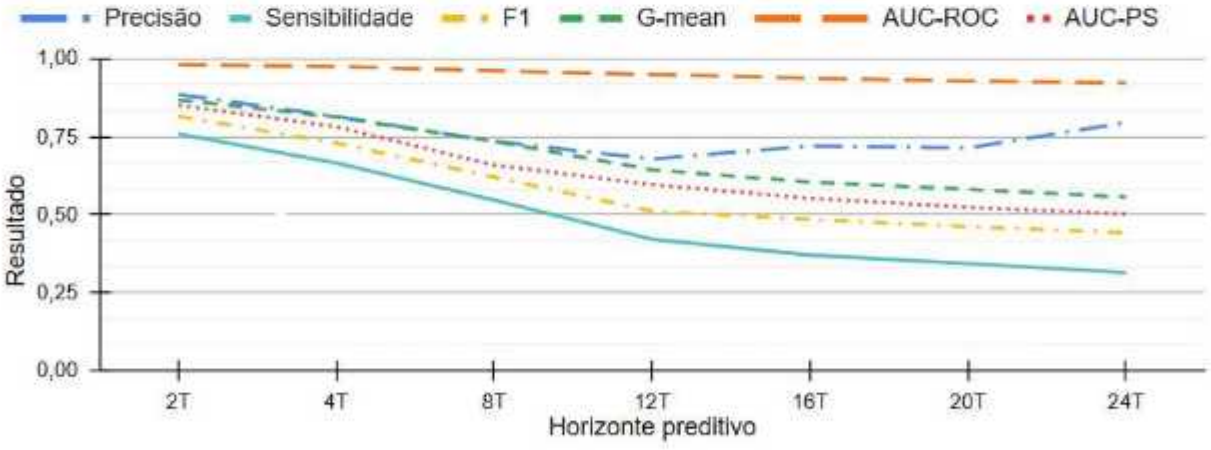
O gráfico de AUC-ROC com aumento do horizonte preditivo se mantém praticamente estável (ver Figura 5.2). O que, segundo Saito & Rehmsmeier (2015) [146], é explicado



(a) Comportamento com reamostragem a 0%.



(b) Comportamento com reamostragem a 50%.



(c) Comportamento com reamostragem a 100%.

Figura 5.2: Gráficos de comportamento do modelo CatBoost usando técnica de reamostragem SMOTE-ENN, variando taxa de balanceamento em 0%, 50% e 100%.

pelo forte desequilíbrio entre as classes. Por isso, eles sugerem a utilização de uma medida adicional, a AUC-PS, para melhor avaliar o desempenho do modelo. Assim, é possível verificar que os resultados obtidos as taxas de 50% e 100% estão próximos de excelente, segundo a escala sugerida por eles (*i.e.* "aleatório", "identificação precoce ruim", "identificação precoce boa", "excelente" e "perfeito"). Onde, "aleatório" é o comportamento de um classificador aleatório e "perfeito" o classificador capaz de identificar precocemente, de mancira correta, todas as empresas do conjunto.

A Precisão difere das demais curvas, pois apresenta uma piora em seus números por causa do aumento da taxa de balanceamento. Quando a taxa de balanceamento é aumentada a curva de Precisão sofre um deslocamento para baixo (piora), enquanto isso, a curva de Sensibilidade sobre um deslocamento para cima (melhora). Além disso, a tendência de queda que ela apresentava sem o balanceamento (Figura 5.2a), após ele, a tendência se inverte, e o horizonte preditivo começa a subir. Com reamostragem a 50%, a inversão de tendência ocorre após o horizonte de 20 trimestres, 20T, saindo de 0,7430 para 0,8127, (Figura 5.2b) e quando é de 100%, a inversão é antecipada, ocorrendo logo após o horizonte de 12T, saindo de 0,6793 para 0,7966 (Figura 5.2c). Indicando que a taxa de balanceamento de 100% supera as anteriores. Portanto, no contexto desta pesquisa, a forma mais adequada de reequilibrar as classes é utilizando SMOTE-ENN a uma taxa de desbalanceamento de 100%.

Como sugerido por Demšar (2006) [163], a confirmação dessa análise foi feita por meio dos testes estatísticos de Friedman [164, 165] e Nemenyi [166]. Primeiramente, foram construídos 6 grupos, um para cada classificador. Cada grupo foi composto pelo classificador sem o uso de reamostragem (1) mais a combinação das técnicas de reamostragem (6) e taxa de balanceamento (2), totalizando 13 combinações. Assim, após os testes de Friedman e Nemenyi foi selecionado a melhor combinação de cada grupo, baseando-se na distância crítica (DC) definida por Nemenyi (1963) [166]. Esses foram classificados entre si a partir das médias da classificação dos resultados por métrica. A tabela Tabela 5.4 apresenta o resultado final dessa análise, organizado em colunas: valor médio da posição de classificação (*ranking*) dos resultados [163] (#), modelo preditivo (Classificador), técnica de reamostragem (Reamostragem), taxa de balanceamento (Tx) e código de referência (Ref.).

A Figura 5.3 apresenta o resultado do teste *post-hoc* de Nemenyi em um gráfico. A linha principal é a régua de posições, sendo a melhor posição representada à direita. A linha mais espessa ligando os classificadores indica que eles não apresentaram diferenças significativas entre si. Porém, o mais importante nessa análise é o fato do classificador C1 está melhor posicionado que os outros, confirmando a análise anterior.

Uma vez definidos o classificador (CatBoost, a técnica de reamostragem SMOTE-ENN

Tabela 5.4: Ordenação de classificadores após testes estatísticos (Friedman e Nemenyi).

#	Classificador	Reamostragem	Tx	Ref.
2,0	CatBoost	SMOTE-ENN	100%	C1
2,5	XGBoost	SMOTE-ENN	100%	C2
3,3	RF	SVM-SMOTE	50%	C3
3,7	AD	SVM-SMOTE	50%	C4
4,3	RL	SMOTE-Tomek	100%	C5
5,2	SVM	SVM-SMOTE	100%	C6

e a taxa de balanceamento de 100%) foi analisado o comportamento do modelo ao longo do tempo, isto é, a medida que os trimestres vão passando. A Figura 5.4 apresenta esse comportamento para os horizontes preditivos de 2, 4 e 8 trimestres (2T, 4T e 8T), que foram escolhidos por apresentarem métricas com resultados acima de 0,5 (ver Figura 5.2c). Os gráficos apresentam em seus eixos verticais os resultados de cada métrica e em seus eixos horizontais os trimestres usados no processo de treinamento por meio da técnica de validação cruzada aninhada para séries temporais [162]. Essa análise é importante para verificar a existência de desvio de conceitos.

Na revisão de literatura feita por Agrahari & Singh (2021) [19] as métricas de Precisão, Sensibilidade, F_1-Score e AUC-ROC podem ser utilizadas para detecção de desvio. Entretanto, neste estudo, devido ao forte desequilíbrio das classes, a curva AUC-ROC apresenta um comportamento praticamente constante evidenciando sua ineficiência para identificar momentos de desvio de conceito. Por outro lado, as demais apresentam instabilidade, evidenciando a existência de alterações nos dados ao longo do tempo, isto é, o desvio de conceito.

Apesar de estarem em patamares diferentes, as curvas de Precisão, Sensibilidade e F_1-Score apresentam comportamento semelhante (ver Figura 5.4). Sendo possível observar picos e vales que evidênciam a existência de desvio de conceito. Entre elas, é possível observar que nem sempre os picos e vales ocorrem no mesmo trimestre ou possuem a mesma duração. Isso ocorre porque quanto maior o horizonte preditivo mais antiga serão as instâncias utilizadas para treinar o modelo, o que tem forte impacto no desvio de conceito.

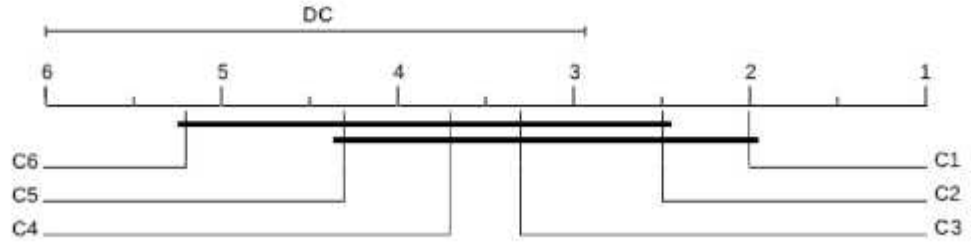
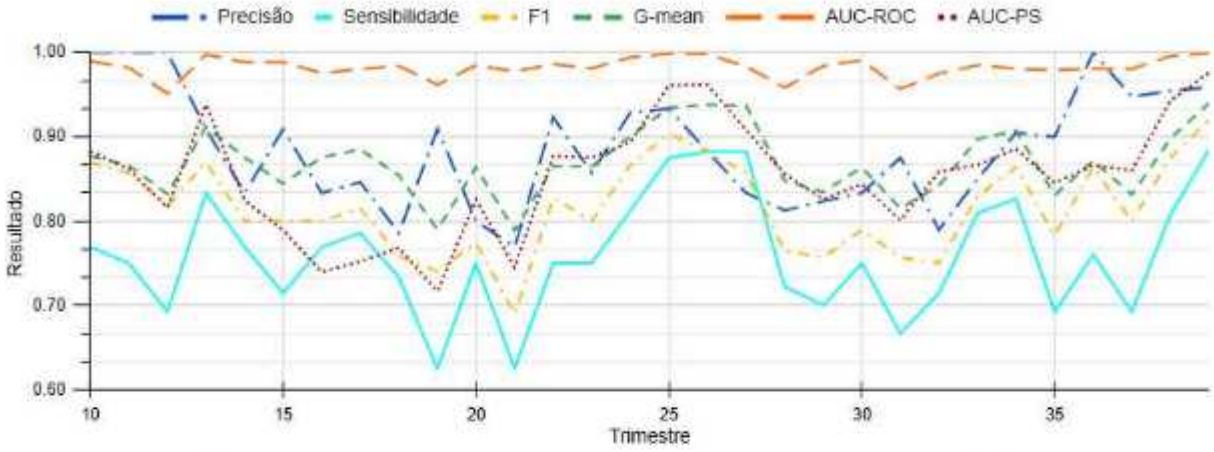
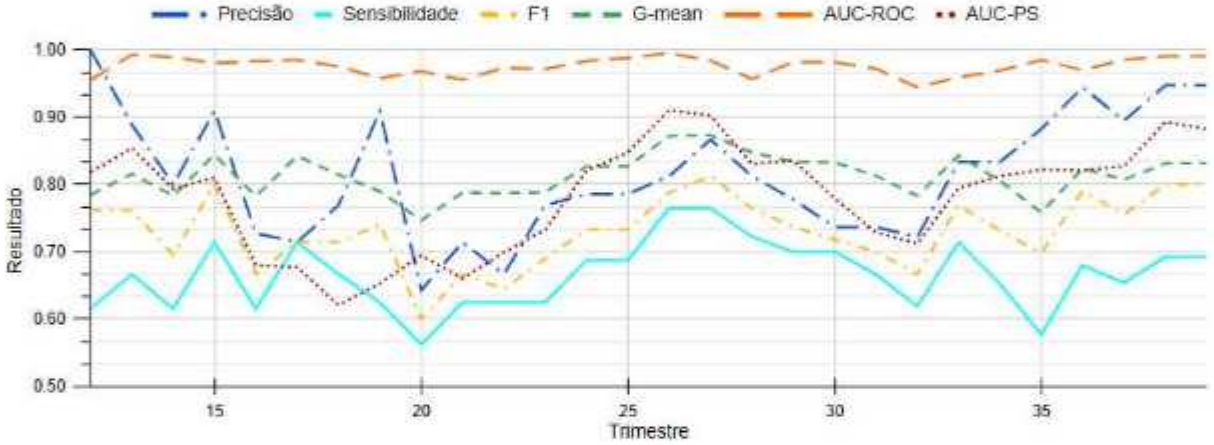


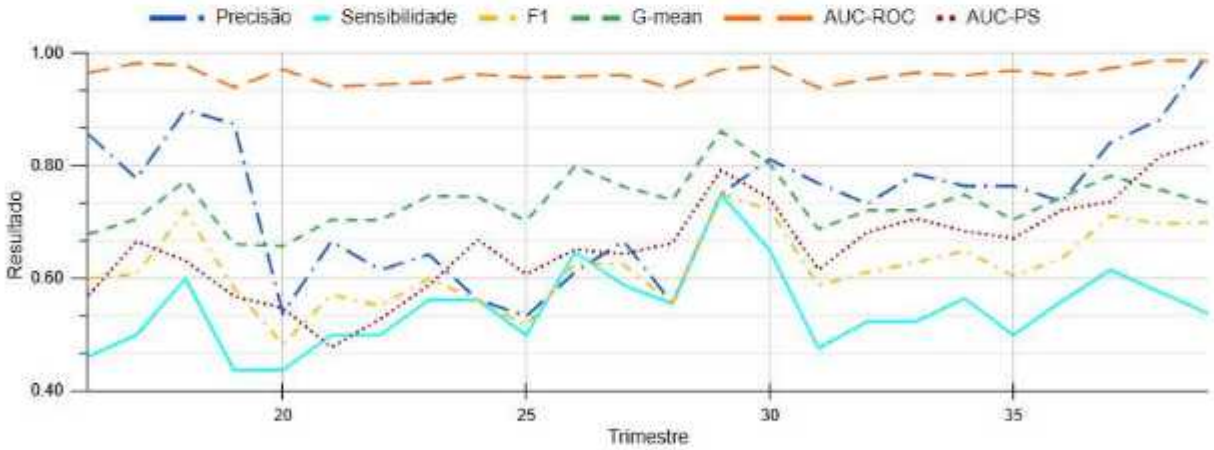
Figura 5.3: Comparação de classificadores com teste de Nemenyi e nível de significância $\alpha = 0,05$. Os classificadores ligados não são significativamente diferentes entre si.



(a) Comportamento com horizonte preditivo de 2 trimestres (2T).



(b) Comportamento com horizonte preditivo de 4 trimestres (4T).



(c) Comportamento com horizonte preditivo de 8 trimestres (8T).

Figura 5.4: Comportamento do CatBoost, com dados balanceados pelo SMOTE-ENN a 100%, ao longo do tempo com variação do horizonte preditivo (2, 4 e 8 trimestres.

Contudo, o comportamento mediano da F_1-Score em relação a Precisão e Sensibilidade é comum em todos os gráficos. Portanto, essa métrica é uma boa referência para análise de desvios de conceito.

Na Figura 5.4a, a curva F_1-Score evidencia algums pontos de mudança nos dados, isto é, desvio de conceito. Primeiramente, no 21º trimestre, observa-se um vale, quando a curva atinge o valor de 0,6897, que também pode ser observado nas curvas de Precisão e Sensibilidade, com valores de 0,7692 e 0,6250, respectivamente. Mais adiante, ocorre uma uma recuperação atingindo o pico no 25º trimestre, com valor de 0,9032. Em seguida, no 26º trimestre, as três curvas se encontram (0,8824) e no trimestre seguinte (27º) a Sensibilidade (0,8824) supera a Precisão (0,8333), voltando a cair no trimestre seguinte. Após isso, observa-se picos e vales menores, porém com uma forte recuperação ao final, ultrapassando o pico anterior com o valor de 0,92.

Na Figura 5.4b, a curva F_1-Score apresenta dois picos e um vale. Após o pico de 0,8 no 15º trimestre, percebe-se uma degradação do desempenho até atingir um vale de 0,6 no 20º trimestre. Posteriormente, ocorreu uma recuperação gradativa, atingindo o novo pico 27º trimestre (0,8125). Por fim, após certa instabilidade, a curva termina sua evolução em um pequeno platô de 2 trimestres (38º e 39º), onde volta a atingir o valor de 0,8. Em relação a Figura 5.4a percebe-se que houve uma suavização na curva, tendência de recuperação das curvas ao final do gráfico e a ocorrência de algum evento no 27º trimestre.

Na Figura 5.4c, neste gráfico a curva F_1-Score apresenta dois picos bem marcados, com posterior baixa nos desempenhos e recuperação subsequente. O primeiro pico ocorre no 18º trimestre (0,72) e a mínima subsequente ocorre no 20º (0,4828). Em seguida, uma recuperação gradativa conduz ao segundo pico, no 29º trimestre (0,75), com subsequente baixa no 31º trimestre (0,5882). Desse ponto em diante, o gráfico caminha para uma recuperação atingindo um platô nos três últimos trimestres (37º, 38º e 39º), com valores de 0,7111, 0,6977 e 0,7. Em relação a Figura 5.4a percebe-se que houve uma suavização na curva e a tendência de recuperação das curvas ao final do gráfico.

Ao observar os três gráficos em conjunto (Figura 5.4), é possível observar uma alteração dos momentos em que são percebidos os picos e vales em cada gráfico, pois a mudança no horizonte preditivo provoca uma mudança nos dados de treinamento. Não obstante as diferenças entre os momentos de ocorrência dos picos e vales, as mudanças nos dados se confirmam em todos eles, caracterizando o desvio de conceito. Além disso, o aumento do horizonte preditivo ocasionou uma suavização das curvas. Especialmente, no processo de recuperação de desempenho que passa a ocorrer de forma mais lenta, quando o horizonte é de 4 e 8 trimestres. A tendência de recuperação em todos eles é observada por meio da análise dos vales, onde o primeiro vale (mais a esquerda) apresenta valor inferior aos valores dos vales seguintes. Essa recuperação pode ser justificada pelo acúmulo de elementos no

histórico, o que faz com que o modelo seja treinado com mais instâncias reais e um pouco menos de instâncias sintéticas.

Comparando os resultados obtidos com outros estudos, observa-se um desempenho equivalente aos melhores resultados de outros estudos. Barboza *et al.* (2021) [40], que utilizou dados desbalanceados e estacionários, obteve para o XGBoost o valor de AUC-ROC igual a 0,9636, enquanto que neste estudo foram obtidos o valores de 0,9424 e 0,9520, utilizando XGBoost e o CatBoost, respectivamente. Shen *et al.* (2020) [44], que utilizadou dados desbalanceados e não estacionários, usando o RF, obteve para as métricas AUC-ROC, F_1-Score e G_{mean} os valores de 0,9138, 0,8003 e 0,8783, respectivamente. Enquanto que, neste estudo o RF obteve, para as mesmas métricas, os valores de 0,9287, 0,3730 e 0,4907. O CatBoost obteve valores de 0,9520, 0,5812 e 0,6865. Sendo observada uma maior diferença nas métricas F_1-Score e G_{mean}, onde este estudo apresentou valores inferiores. O melhor desempenho apresentado no estudo de Shen ocorre devido a taxa de desbalanceamento, pois neste estudo a classe minoritária representa 2,73% do total, enquanto que no estudo de Shen ela representa 30%.

Capítulo 6

Conclusão

Diante de ambientes acompanhados em tempo real ou cada vez mais monitorados, este estudo buscou identificar formas de monitorar a condição financeira de empresas e identificar as situações de DF. Essa capacidade é de grande importância para os agentes de mercado, como investidores, instituições financeiras fornecedoras de empréstimos e agentes de governos responsáveis por monitorar o mercado financeiro, como o Banco Central do Brasil. Dessa forma, por meio de modelos de IA, este estudo buscou classificar as empresas listadas na bolsa de valores brasileira (B3) em SDF e em DF, considerando essas informações como não estacionárias. Devido a ausência de bases de dados não estacionárias para classificação de empresas em DF, disponíveis para acesso público, este estudo elaborou uma base de dados de indicadores econômico-financeiros extraídos de informações contábeis do Portal de Dados Abertos da CVM.

A base de dados produzida contém 84 indicadores de 915 empresas distintas em um período de 10 anos, de 2011 a 2020, organizados em 40 trimestres para processamento de forma não estacionária. Portanto, toda a base de dados tem 23.834 registros, sendo 23.183 registros de empresas SDF e 651 registros de empresas em DF, representado em percentuais como 97,27% e 2,73%, respectivamente. Devido ao forte desbalanceamento das classes, foi necessário a utilização de técnicas de reamostragem para evitar o viés da classe majoritária. Este estudo identificou que a técnica de reamostragem SMOTE é uma das mais populares, com vastos trabalhos sobre ela e APIs para produção de novos estudos. Por isso, neste estudo foi utilizada a técnica de SMOTE e algumas de suas variantes a B-SMOTE, ADASYN, SVM-SMOTE, SMOTE-ENN e SMOTE-Tomek. Dessa forma, foram produzidas instâncias sintéticas para reequilibrar as classes na proporção de 0%, 50% e 100%.

A revisão sistemática de literatura conduzida nesse estudo permitiu identificar o estado da arte na área de predição de DF, indicando os modelos de IA como superiores aos modelos estatísticos. Alguns dos modelos mais citados para o tratamento de bases estaci-

onárias e não estacionárias foram RL, SVM ,RF, AD, XGBoost e CatBoost. Entretanto, estudos com bases não estacionárias ainda são poucos e fizeram uso de modelos semelhantes [44, 43]. Por isso, esses modelos foram selecionados para este estudo. Todavia, o uso desses modelos preditivos com dados não estacionários requer cuidado. Por isso, no treinamento dos modelos foi utilizada a técnica de validação cruzada aninhada para séries temporais. Porém, esse não é o único desafio, uma vez que os dados evoluem com o tempo e apresentam desvio de conceito, o desempenho dos modelos tendem a degradar ao longo do tempo.

Entre as várias técnicas para identificar, tratar e minimizar o desvio de conceito, optou-se por utilizar a técnica de janela deslizante em conjunto com um histórico que utiliza um mecanismo de esquecimento. Essa escolha foi motivada pela simplicidade, eficiência e popularidade da técnica, que também foi empregada por Shen *et al.* (2020) [44]. O estudo não buscou aprofundar a pesquisa sobre os diferentes tipos de desvio de conceito presentes no conjunto de dados, nem investigar as motivações por trás deles ou variar os parâmetros de configuração da técnica adotada. Portanto, a janela deslizante utilizada foi definida com um tamanho de 8 trimestres, o histórico de tamanho livre e o mecanismo de esquecimento foi configurado com um coeficiente igual a 1. Esses valores permaneceram constantes em todos os experimentos.

No total, foram realizados 546 experimentos, combinando modelos preditivos, técnicas de reamostragem, taxa de balanceamento e horizonte preditivo. Observou-se que as métricas mais frequentemente utilizadas foram Precisão, Sensibilidade, F_1-Score, G_{mean} e AUC-ROC, devido à natureza desbalanceada dos dados. Este estudo adicionou a métrica AUC-PS, sugerida por Saito & Rehmsmeier (2015) [146], para análise de modelos em dados desbalanceados, o que ainda não havia sido explorado em estudos de previsão de DF em ambientes de fluxo de dados desbalanceados. Com o uso dessas métricas, foram gerados 3.276 resultados de experimentos, que foram analisados empiricamente por meio de raciocínio dedutivo e confirmados por testes estatísticos, como o teste de Friedman e o teste *post-hoc* de Nemenyi. Concluiu-se que, em geral, a combinação mais eficiente utilizou a técnica de reamostragem SMOTE-ENN, com taxa de balanceamento de 100%, e o modelo preditivo CatBoost, obtendo resultados similares aos de Shen *et al.* (2020) [44], mas em um conjunto de dados com um maior fator de desbalanceamento.

Este estudo demonstra a viabilidade preditiva do uso de indicadores econômico-financeiros processados como fluxo de dados organizados por trimestres para identificar empresas em estado de DF, mesmo em cenários com dados fortemente desbalanceados. Em particular, a utilização da combinação mais eficiente mostra que a reamostragem é capaz de melhorar os resultados de todos os modelos avaliados, incluindo aqueles considerados menos suscetíveis ao desbalanceamento, como RF, XGBoost e CatBoost. Além disso, observou-se

que, apesar do impacto negativo causado pelo aumento do horizonte preditivo nos resultados, a reamostragem é capaz de minimizar esses impactos, permitindo predições com maior antecedência. No contexto de fluxo de dados, constatou-se que a combinação escolhida possui capacidade de recuperação após um desvio de conceito, e que o aumento do horizonte preditivo é inversamente proporcional à velocidade de recuperação após tal desvio.

Além disso, este estudo contribui com recomendações objetivas, tais como: o uso de técnicas de reamostragem em fluxo de dados desbalanceados, como o SMOTE-ENN, devido à sua capacidade de melhorar os resultados obtidos; cautela ao aumentar o horizonte preditivo na predição de DF, devido ao impacto significativo que essa medida pode ter; considerar o CatBoost como uma opção adequada para a classificação binária em fluxo de dados desbalanceados; empregar técnicas de tratamento de desvio de conceito, como a janela deslizante, histórico e mecanismo de esquecimento, para desenvolver soluções mais autônomas capazes de se recuperar de perda de desempenho causada por mudanças nos dados; utilizar a métrica AUC-PS como complemento à métrica AUC-ROC, uma vez que esta última, isoladamente, não é capaz de medir adequadamente o desempenho do modelo ao longo do tempo. Por fim, recomenda-se realizar a predição de DF de forma não estacionária, levando em consideração que, em situações reais, os indicadores econômico-financeiros das empresas são suscetíveis a desvios de conceito [43].

Alguns parâmetros deste estudo foram fixados para permitir uma melhor comparação com outros estudos, como o de Shen et al. (2020) [44], que utilizou o mesmo coeficiente de esquecimento igual a 1. No entanto, após verificar a viabilidade dos métodos empregados, estudos futuros podem variar esses valores para encontrar os mais adequados para o caso específico. O aumento do tamanho da janela deslizante também pode ser testado, pois isso introduziria um maior número de instâncias para treinamento do modelo, com potencial para melhorar o desempenho. Outra opção é explorar a janela deslizante adaptável [142]. Neste momento, as informações dos anos de 2021 e 2022 já estão disponíveis no portal de dados abertos da CVM. Incorporar essas informações aumentaria o conjunto de dados e impactaria o histórico. Com o coeficiente de esquecimento adequado, seria possível reduzir a necessidade de instâncias sintéticas da classe minoritária. Além disso, mais pesquisas poderiam ser feitas sobre o desvio de conceito, a fim de identificar diferentes tipos de desvios e adaptar os modelos após a detecção [19]. Sistemas de recomendação e agentes autônomos podem ser incrementados para fornecer informações sobre investimentos. Por fim, para contribuir com estudos futuros, a base de dados utilizada neste estudo está disponível no Github[1], juntamente com o código fonte para sua construção. Isso facilita a atualização com dados provenientes da CVM e permite a automação desse processo.

[1] `https://github.com/rubensmchaves/ml-fdp`

Referências

[1] Gomes, Heitor Murilo, Jesse Read, Albert Bifet, Jean Paul Barddal e João Gama: *Machine learning for streaming data: State of the art, challenges, and opportunities.* ACM SIGKDD Exploration Newsletter, 21(2):6—-22, 2019. 1, 10, 24, 34

[2] He, Haibo, Sheng Chen, Kang Li e Xin Xu: *Incremental learning from stream data.* IEEE Transactions on Neural Networks, 22(12):1901–1914, 2011. 1, 24

[3] Altman, Edward I.: *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy.* The Journal of Finance, 23(4):589–609, 1968. 1, 2, 9, 23, 38

[4] Korol, Tomasz: *Dynamic bankruptcy prediction models for european enterprises.* Journal of Risk and Financial Management, 12(4), 2019. 1, 2, 19

[5] Santos Fernandes, Maria Helena dos: *Instrução normativa cvm nº 480.* Diário Oficial da União, 1(235):28–36, 2009. 1, 24, 45

[6] Lu, Jie, Anjin Liu, Fan Dong, Feng Gu, João Gama e Guangquan Zhang: *Learning under concept drift: A review.* IEEE Transactions on Knowledge and Data Engineering, 31(12):2346–2363, 2019. 1, 2

[7] Barboza, Flavio, Herbert Kimura e Edward Altman: *Machine learning models and bankruptcy prediction.* Expert Systems with Applications, 83:405–417, 2017. 1, 2, 5, 9, 17, 24, 38

[8] Alaka, Hafiz A., Lukumon O. Oyedele, Hakeem A. Owolabi, Vikas Kumar, Saheed O. Ajayi, Olugbenga O. Akinade e Muhammad Bilal: *Systematic review of bankruptcy prediction models: Towards a framework for tool selection.* Expert Systems with Applications, 94:164–184, 2018. 2, 4, 19, 38, 39

[9] Balcaen, Sofie e Hubert Ooghe: *35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems.* The British Accounting Review, 38(1):63–93, 2006. 2

[10] Shi, Yin e Xiaoni Li: *An overview of bankruptcy prediction models for corporate firms: A systematic literature review.* Intangible Capital, 15:114, 2019. 2

[11] Oussous, Ahmed, Fatima Zahra Benjelloun, Ayoub Ait Lahcen e Samir Belfkih: *Big data technologies: A survey.* Journal of King Saud University - Computer and Information Sciences, 30(4):431–448, 2018. 2, 27

[12] Gantz, John e David Reinsel: *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east.* Em *IDC IVIEW*, 2012. 2

[13] Sleeman IV, William C. e Bartosz Krawczyk: *Multi-class imbalanced big data classification on spark.* Knowledge-Based Systems, 212:106598, 2021. 2, 19

[14] Witten, Ian H., Eibe Frank e Mark A. Hall: *Chapter 10 - Introduction to Weka.* The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, third edition edição, 2011. 2

[15] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot e Édouard Duchesnay: *Scikit-learn: Machine learning in python.* The Journal of Machine Learning Research, 12:2825—-2830, 2011. 2

[16] Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu e Xiaoqiang Zheng: *Tensorflow: Large-scale machine learning on heterogeneous distributed systems.* Em *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.* arXiv, 2016. 2

[17] Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai e Soumith Chintala: *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* Curran Associates Inc., 2019. 2

[18] Gama, João, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy e Abdelhamid Bouchachia: *A survey on concept drift adaptation.* ACM Computing Surveys, 46(44):1–37, 2014. 2, 3, 26, 48

[19] Agrahari, Supriya e Anil Kumar Singh: *Concept drift detection in data stream mining : A literature review.* Journal of King Saud University - Computer and Information Sciences, 2021. 2, 3, 24, 25, 26, 27, 29, 31, 34, 63, 69

[20] Ditzler, Gregory, Manuel Roveri, Cesare Alippi e Robi Polikar: *Learning in non-stationary environments: A survey.* IEEE Computational Intelligence Magazine, 10(4):12–25, 2015. 2, 3, 31

[21] Schlimmer, Jeffrey C. e Richard H. Granger: *Incremental learning from noisy data.* Machine Learning, 1(3):317—-354, 1986. 3

[22] Widmer, Gerhard e Miroslav Kubat: *Learning in the presence of concept drift and hidden contexts.* Machine Learning, 23(1):69—-101, 1996. 3

[23] Tsymbal, Alexey: *The problem of concept drift: definitions and related work.* Computer Science Department, Trinity College Dublin, 106(2):58, 2004. 3

[24] Li, Zeng, Wenchao Huang, Yan Xiong, Siqi Ren e Tuanfei Zhu: *Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm.* Knowledge-Based Systems, 195:105694, 2020. 4, 5, 19, 20, 41

[25] Silva, Thiago Christiano, Michel da Silva Alexandre e Benjamin Miranda Tabak: *Bank lending and systemic risk: A financial-real sector network approach with feedback.* Journal of Financial Stability, 38:98–118, 2017. 4, 21, 22, 25

[26] Fernández, Alberto, Salvador García, Mikel Galar, Ronaldo Prati, Bartosz Krawczyk e Francisco Herrera: *Learning from Imbalanced Data Sets.* Springer Cham, 2018. 4, 24, 25, 26, 27, 34, 41, 42, 43, 48

[27] Fernández, Alberto, Salvador García, Francisco Herrera e Nitesh V. Chawla: *Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary.* Journal of Artificial Intelligence Research, 61(1):863—-905, 2018. 4

[28] Wang, Shuo, Leandro L. Minku e Xin Yao: *A systematic study of online class imbalance learning with concept drift.* IEEE Transactions on Neural Networks and Learning Systems, 29(10):4802–4821, 2018. 4, 10

[29] Pisani, Paulo Henrique e Ana Carolina Lorena: *A systematic review on keystroke dynamics.* Journal of the Brazilian Computer Society, 19(4):573–587, 2013. 4, 8

[30] Biolchini, Jorge, P Gomes Mian, A Candida Cruz Natali e G Horta Travassos: *Systematic review in software engineering.* System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES, 679(05), 2005. 4

[31] Clement, Claudiu: *Machine learning in bankruptcy prediction - a review.* Journal of Public Administration, Finance and Law, 17:178–197, 2020. 4, 15, 17, 18, 39

[32] Duarte, Denize Lemos e Flávio Luiz de Moraes Barboza: *Forecasting financial distress with machine learning – a review.* Future Studies Research Journal: Trends and Strategies, 12(3):528—-574, 2020. 4, 9

[33] Demyanyk, Yuliya e Otto Van Hemert: *Understanding the subprime mortgage crisis.* The Review of financial studies, 24(6):1848–1880, 2011. 5

[34] Castro, Paulo de Tarso Amorim: *Desastres de mariana e brumadinho.* Caderno de geografia (Belo Horizonte, Brazil), 31(1):196, 2021. 5

[35] Chowdhury, Emon Kalyan, Iffat Ishrat Khan e Bablu Kumar Dhar: *Catastrophic impact of covid-19 on the global stock markets and economic activities.* Business and Society Review, 2021. 5, 22

[36] Brereton, Pearl, Barbara A. Kitchenham, David Budgen, Mark Turner e Mohamed Khalil: *Lessons from applying the systematic literature review process within the software engineering domain.* Journal of Systems and Software, 80(4):571—583, 2007. 8

[37] Budgen, David e Pearl Brereton: *Performing systematic literature reviews in software engineering.* Em *Proceedings of the 28th International Conference on Software Engineering*, página 1051–1052, 2006. 8

[38] Alam, Talha Mahboob, Kamran Shaukat, Mubbashar Mushtaq, Yasir Ali, Matloob Khushi, Suhuai Luo e Abdul Wahab: *Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World.* The Computer Journal, 64(11):1731–1746, 2020. 9, 11, 15, 19, 23, 24, 40

[39] Shi, Yin e Xiaoni Li: *A bibliometric study on intelligent techniques of bankruptcy prediction for corporate firms.* Heliyon, 5(12):12, 2019. 9

[40] Barboza, Flávio Luiz de Moraes, Denize Lemos Duarte e Michele Aparecida Cunha: *Anticipating corporate's distresses.* EXACTA Engenharia de Produção, 20(2), 2022. 15, 17, 18, 19, 24, 39, 66

[41] Wang, Haoming e Xiangdong Liu: *Undersampling bankruptcy prediction: Taiwan bankruptcy data.* PLOS ONE, 16(7):1–17, 2021. 15, 40

[42] Sun, Jie, Hamido Fujita, Peng Chen e Hui Li: *Dynamic financial distress prediction with concept drift based on time weighting combined with adaboost support vector machine ensemble.* Knowledge-Based Systems, 120:4–14, 2017. 15, 39

[43] Sun, Jie, Mengjie Zhou, Wenguo Ai e Hui Li: *Dynamic prediction of relative financial distress based on imbalanced data stream: from the view of one industry.* Risk Management, 21(4):215–242, 2019. 15, 68, 69

[44] Shen, Feng, Yongyong Liu, Run Wang e Wei Zhou: *A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment.* Knowledge-Based Systems, 192:105365, 2020. 15, 18, 19, 20, 32, 35, 39, 40, 41, 48, 66, 68, 69

[45] Bragoli, Daniela, Camilla Ferretti, Piero Ganugi, Giovanni Marseguerra, Davide Mezzogori e Francesco Zammori: *Machine-learning models for bankruptcy prediction: do industrial variables matter?* Spatial Economic Analysis, 17(2):156–177, 2022. 18

[46] Zou, Yao, Changchun Gao e Han Gao: *Business failure prediction based on a cost-sensitive extreme gradient boosting machine.* IEEE Access, 10:42623–42639, 2022. 18

[47] Chen, Ying, Jifeng Guo, Junqin Huang e Bin Lin: *A novel method for financial distress prediction based on sparse neural networks with $l_{1/2}$ regularization.* International Journal of Machine Learning and Cybernetics, 13(7):2089–2103, 2022. 18

[48] Succurro, Marianna: *Financial bankruptcy across european countries.* International Journal of Economics and Finance, 9(7):132–146, 2017. 18

[49] Pilch, Bartłomiej: *An analysis of the effectiveness of bankruptcy prediction models – an industry approach.* Folia Oeconomica Stetinensia, 21(2):76–96, 2021. 18

[50] Tomczak, Sebastian: *Polish companies bankruptcy data.* UCI Machine Learning Repository, 2016. DOI: `https://doi.org/10.24432/C5F600`. 18

[51] Liang, Deron e Chih Fong Tsai: *Taiwanese bankruptcy prediction.* UCI Machine Learning Repository, 2020. `https://archive-beta.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction`. 18

[52] Lombardo, Gianfranco, Mattia Pellegrino, George Adosoglou, Stefano Cagnoni, Panos M. Pardalos e Agostino Poggi: *Machine learning for bankruptcy prediction in the american stock market: Dataset and benchmarks.* Future Internet, 14(8), 2022. 18

[53] Liang, Deron, Chia Chi Lu, Chih Fong Tsai e Guan An Shih: *Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study.* European Journal of Operational Research, 252(2):561–572, 2016. 18, 22, 40

[54] Tobback, Ellen, Tony Bellotti, Julie Moeyersoms, Marija Stankova e David Martens: *Bankruptcy prediction for smes using relational data.* Decision Support Systems, 102:69–81, 2017. 18

[55] Mai, Feng, Shaonan Tian, Chihoon Lee e Ling Ma: *Deep learning models for bankruptcy prediction using textual disclosures.* European Journal of Operational Research, 274(2):743–758, 2019. 18

[56] Lukason, Oliver e Art Andresson: *Tax Arrears Versus Financial Ratios in Bankruptcy Prediction.* JRFM, 12(4):1–13, 2019. 18

[57] Liu, Weike, Hang Zhang, Zhaoyun Ding, Qingbao Liu e Cheng Zhu: *A comprehensive active learning method for multiclass imbalanced data streams with concept drift.* Knowledge-Based Systems, 215:106778, 2021. 19, 20, 41

[58] Sun, Yange, Meng Li, Lei Li, Han Shao e Yi Sun: *Cost-sensitive classification for evolving data streams with concept drift and class imbalance.* Computational Intelligence and Neuroscience, 2021. 19

[59] Wegier, Weronika e Pawel Ksieniewicz: *Application of imbalanced data classification quality metrics as weighting methods of the ensemble data stream classification algorithms.* Entropy (Basel, Switzerland), 22(8):849, 2020. 19

[60] Ghazikhani, Adel, Reza Monsefi e Hadi Sadoghi Yazdi: *Ensemble of online neural networks for non-stationary and imbalanced data streams.* Neurocomputing, 122:535–544, 2013. 19, 20, 41

[61] Wang, Shuo, Leandro L. Minku e Xin Yao: *Resampling-based ensemble methods for online class imbalance learning.* IEEE Transactions on Knowledge and Data Engineering, 27(5):1356–1368, 2015. 19

[62] Migućis, Vera L., Ana S. Camanho e José Borges: *Predicting direct marketing response in banking: comparison of class imbalance methods.* Service Business, 11(4):831–849, 2017. 19

[63] Ditzler, Gregory, Robi Polikar e Nitesh Chawla: *An incremental learning algorithm for non-stationary environments and class imbalance.* Em *2010 20th International Conference on Pattern Recognition*, páginas 2997–3000, 2014. 19, 26

[64] Chen, Sheng e Haibo He: *Towards incremental learning of nonstationary imbalanced data stream: a multiple selectively recursive approach.* Evolving Systems, 2(1):35–50, 2011. 19, 20, 41

[65] Toor, Affan Ahmed, Muhammad Usman, Farah Younas, Alvis Cheuk M. Fong, Sajid Ali Khan e Simon Fong: *Mining massive e-health data streams for iomt enabled healthcare systems.* Sensors, 20(7), 2020. 19

[66] Setiawan, Budi Darma, Uwe Serdült e Victor Kryssanov: *A machine learning framework for balancing training sets of sensor sequential data streams.* Sensors, 21(20), 2021. 19

[67] Wang, Jia Bao, Chun An Zou e Guang Hui Fu: *Awsmote: An svm-based adaptive weighted smote for class-imbalance learning.* Scientific Programming, 2021, 2021. 19

[68] Li, Qiude, Qingyu Xiong, Shengfen Ji, Yang Yu, Chao Wu e Min Gao: *Incremental semi-supervised extreme learning machine for mixed data stream classification.* Expert Systems with Applications, 185:115591, 2021. 19

[69] Zięba, Maciej, Sebastian K. Tomczak e Jakub M. Tomczak: *Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction.* Expert Systems with Applications, 58:93–101, 2016. 19, 20, 40

[70] Alaminos, David, Agustín del Castillo e Manuel Ángel Fernández: *A global model for bankruptcy prediction.* PLOS ONE, 11(11):1–18, 2016. 19, 24

[71] Montiel, Jacob, Jesse Read, Albert Bifet e Talel Abdessalem: *Scikit-multiflow: A multi-output streaming framework.* Journal of Machine Learning Research, 19(72):1–5, 2018. 20

[72] Ferguson, Niall: *A ascensão do dinheiro.* Crítica, 3ª edição, 2020. 21

[73] Felloni, Giuseppe: *A profile of genoa's 'casa di san giorgio' (1407-1805): A turning point in the history of credit.* Rivista di storia economica, Italian Review of Economic History, 3:335–346, 2010. 21

[74] Tabak, Benjamin M., Ana Clara Noronha e Daniel Cajueiro: *Bank capital buffers, lending growth and economic cycle: empirical evidence for brazil.* Em *2nd BIS Consultative Council for the Americas (2nd BIS CCA)*, 2011. 21

[75] Kashyap, Anil K. e Jeremy C. Stein: *The role of banks in monetary policy: a survey with implications for the european monetary union.* Economic Perspectives, XXI(5):2–18, 1997. 21

[76] Duarte, Fernando e Collin Jones: *Empirical network contagion for u.s. financial institutions.* FRB of NY Staff Report, 1(826), 2017. 22

[77] Eichengreen, Barry, Ashoka Mody, Milan Nedeljkovic e Lucio Sarno: *How the subprime crisis went global: Evidence from bank credit default swap spreads.* Journal of International Money and Finance, 31(5):1299–1318, 2012. 22

[78] Ohanian, Lee E.: *What – or who – started the great depression?* Journal of Economic Theory, 144(6):2310–2335, 2009. 22

[79] Aremu, Johnson Olaosebikan: *A historical analysis of the nature, causes and impact of the foreign debt crisis in latin america, 1970- 1980.* Humanities and Social Sciences Letters, 6(3):74—-83, 2018. 22

[80] Stuchlikova, Zuzana: *Japan´s lost decade: on the development of the japanese economy in the 1990s.* Journal of International Relations, 10:129–152, 2012. 22

[81] Mishkin, Frederic S.: *Lessons from the tequila crisis.* Journal of Banking & Finance, 23(10):1521–1533, 1999. 22

[82] Duygan-Bump, Burcu, Alexey Levkov e Judit Montoriol-Garriga: *Financing constraints and unemployment: Evidence from the great recession.* Journal of Monetary Economics, 75:89–105, 2015. 22

[83] Varotto, Simone e Lei Zhao: *Systemic risk and bank size.* Journal of International Money and Finance, 82:45–70, 2018. 22

[84] Merwin, Charles L.: *Financing Small Corporations in Five Manufacturing Industries, 1926–36.* National Bureau of Economic Research, Inc, 1942. 23

[85] Martorano, Luca: *Company bankruptcy prediction.* Kaggle, 2021. `https://www.kaggle.com/code/marto24/bankruptcy-detection`, Accessed: 2022-05-21. 23

[86] Comissão de Valores Monetários: *Resolução CVM Nº 155, de 23 de Junho de 2022.* Diário Oficial da União, 2022. `https://conteudo.cvm.gov.br/legislacao/resolucoes/resol155.html`, acesso em 2022-06-23. 24

[87] Hosaka, Tadaaki: *Bankruptcy prediction using imaged financial ratios and convolutional neural networks.* Expert Systems with Applications, 117:287–299, 2019. 24

[88] Jain, Sanjay, Steffen Lange e Sandra Zilles: *Towards a better understanding of incremental learning.* Em *Algorithmic Learning Theory*, páginas 169–183, 2006. 24, 25, 26

[89] Žliobaitė, Indrė: *Learning under concept drift: an overview.* Em *Cornell University: arXiv*, 2010. 24

[90] Hoens, T. Ryan, Robi Polikar e Nitesh V. Chawla: *Learning from streaming data with concept drift and imbalance: an overview.* Progress in Artificial Intelligence, 1(1):89–101, 2012. 26, 48

[91] Demšar, Jaka e Zoran Bosnić: *Detecting concept drift in data streams using model explanation.* Expert Systems with Applications, 92:546–559, 2018. 26, 30

[92] Bu, Li, Dongbin Zhao e Cesare Alippi: *An incremental change detection test based on density difference estimation.* IEEE Transactions on Systems, Man, and Cybernetics: Systems, 47(10):2714–2726, 2017. 26, 29

[93] Frías-Blanco, Isvani, José del Campo-Ávila, Gonzalo Ramos-Jiménez, Rafael Morales-Bueno, Agustín Ortiz-Díaz e Yailé Caballero-Mota: *Online and non-parametric drift detection methods based on hoeffding's bounds.* IEEE Transactions on Knowledge and Data Engineering, 27(3):810–823, 2015. 26, 27

[94] Liu, Dong, YouXi Wu e He Jiang: *Fp-elm: An online sequential learning algorithm for dealing with concept drift.* Neurocomputing, 207:322–334, 2016. 26, 30

[95] Gama, João, Pedro Medas, Gladys Castillo e Pedro Rodrigues: *Learning with drift detection.* Em *Advances in Artificial Intelligence – SBIA 2004*, páginas 286–295, 2004. 27

[96] Baena-García, Manuel, José Campo-Ávila, Raúl Fidalgo-Merino, Albert Bifet, Ricard Gavald e Rafael Morales-Bueno: *Early drift detection method.* In Fourth international workshop on knowledge discovery from data streams, 6:77—-86, 2006. 27

[97] Ross, Gordon J., Niall M. Adams, Dimitris K. Tasoulis e David J. Hand: *Exponentially weighted moving average charts for detecting concept drift.* Pattern Recognition Letters, 33(2):191–198, 2012. 27

[98] Barros, Roberto S.M., Danilo R.L. Cabral, Paulo M. Gonçalves e Silas G.T.C. Santos: *Rddm: Reactive drift detection method.* Expert Systems with Applications, 90:344–355, 2017. 27

[99] Gama, Joãoo e Gladys Castillo: *Learning with local drift detection.* Em *Advanced Data Mining and Applications*, páginas 42–55, 2006. 27

[100] Xu, Shuliang e Junhong Wang: *Dynamic extreme learning machine for data stream classification.* Neurocomputing, 238:433–449, 2017. 27

[101] Wang, Pingfan, Nanlin Jin e Gerhard Fehringer: *Concept drift detection with false positive rate for multi-label classification in iot data stream.* Em *2020 International Conference on UK-China Emerging Technologies (UCET)*, páginas 1–4, 2020. 27

[102] Nishida, Kyosuke e Koichiro Yamauchi: *Detecting concept drift using statistical testing.* Em Corruble, Vincent, Masayuki Takeda e Einoshin Suzuki (editores): *Discovery Science*, páginas 264–269, 2007. 27, 29

[103] Song, Ge, Yunming Ye, Haijun Zhang, Xiaofei Xu, Raymond Y.K. Lau e Feng Liu: *Dynamic clustering forest: An ensemble framework to efficiently classify textual data stream with concept drift.* Information Sciences, 357:125–143, 2016. 27

[104] Cabral, Danilo Rafael de Lima e Roberto Souto Maior de Barros: *Concept drift detection based on fishers exact test.* Inf. Sci., 442(C):220—-234, 2018. 27

[105] de Mello, Rodrigo F., Yule Vaz, Carlos H. Grossi e Albert Bifet: *On learning guarantees to unsupervised concept drift detection on data streams.* Expert Systems with Applications, 117:90–102, 2019, ISSN 0957-4174. 27, 29

[106] Bifet, Albert e Ricard Gavaldà: *Learning from time-changing data with adaptive windowing.* Em *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)*, páginas 443–448, 2007. 29

[107] Bifet, Albert: *Adaptive learning and mining for data streams and frequent patterns.* SIGKDD Explor. Newsl., 11(1):55—-56, 2009. 29

[108] Huang, David Tse Jung, Yun Sing Koh, Gillian Dobbie e Russel Pears: *Detecting volatility shift in data streams.* Em *2014 IEEE International Conference on Data Mining*, páginas 863–868, 2014. 29

[109] Gözüaçık, Ömer e Fazli Can: *Concept learning using one-class classifiers for implicit drift detection in evolving data streams.* Artificial Intelligence Review, 54(5):3725–3747, 2021. 29

[110] Li, Peipei, Xindong Wu, Xuegang Hu e Hao Wang: *Learning concept-drifting data streams with random ensemble decision trees.* Neurocomputing, 166:68–83, 2015. 29

[111] Pesaranghader, Ali e Herna L. Viktor: *Fast hoeffding drift detection method for evolving data streams.* Em *Machine Learning and Knowledge Discovery in Databases*, páginas 96–111, 2016. 29

[112] Barros, Roberto Souto Maior de, Juan Isidro González Hidalgo e Danilo Rafael de Lima Cabral: *Wilcoxon rank sum test drift detector.* Neurocomputing, 275:1954–1963, 2018. 29

[113] Bach, Stephen H. e Marcus A. Maloof: *Paired learners for concept drift.* Em *2008 Eighth IEEE International Conference on Data Mining*, páginas 23–32, 2008. 29

[114] Page, E. S.: *Continuous inspection schemes.* Biometrika, 41(1/2):100–115, 1954. 29, 30

[115] Maciel, Bruno Iran Ferreira, Silas Garrido Teixeira Carvalho Santos e Roberto Souto Maior Barros: *A lightweight concept drift detection ensemble.* Em *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, páginas 1061–1068, 2015. 29

[116] Yu, Shujian, Xiaoyang Wang e José C. Príncipe: *Request-and-reverify: Hierarchical hypothesis testing for concept drift detection with expensive labels.* Em *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, páginas 3033–3039. International Joint Conferences on Artificial Intelligence Organization, julho 2018. 29

[117] Raza, Haider, Girijesh Prasad e Yuhua Li: *Ewma model based shift-detection methods for detecting covariate shifts in non-stationary environments.* Pattern Recognition, 48(3):659–669, 2015. 29

[118] Yu, Shujian, Zubin Abraham, Heng Wang, Mohak Shah, Yantao Wei e José C. Príncipe: *Concept drift detection and adaptation with hierarchical hypothesis testing.* Journal of the Franklin Institute, 356(5):3187–3215, 2019. 29

[119] Park, Jin Man e Jong Hwan Kim: *Online recurrent extreme learning machine and its application to time-series prediction.* Em *2017 International Joint Conference on Neural Networks (IJCNN)*, páginas 1983–1990, 2017. 29

[120] Khezri, Shirin, Jafar Tanha, Ali Ahmadi e Arash Sharifi: *Stds: self-training data streams for mining limited labeled data in non-stationary environment.* Applied Intelligence, 50(5):1448–1467, 2020. 29

[121] Qahtan, Abdulhakim A., Basma Alharbi, Suojin Wang e Xiangliang Zhang: *A pca-based change detection framework for multidimensional data streams: Change detection in multidimensional data streams.* Em *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 935—-944, 2015. 29

[122] Song, Xiuyao, Mingxi Wu, Christopher Jermaine e Sanjay Ranka: *Statistical change detection for multi-dimensional data.* Em *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 667—-676, 2007. 29

[123] Bu, Li, Cesare Alippi e Dongbin Zhao: *Ensemble lsdd-based change detection tests.* Em *2016 International Joint Conference on Neural Networks (IJCNN)*, páginas 4064–4069, 2016. 29

[124] Liu, A, Y Song, G Zhang e J Lu: *Regional concept drift detection and density synchronized drift adaptation.* IJCAI International Joint Conference on Artificial Intelligence, páginas 2280–2286, 2017. 29

[125] Gu, Feng, Guangquan Zhang, Jie Lu e Chin Teng Lin: *Concept drift detection based on equal density estimation.* Em *2016 International Joint Conference on Neural Networks (IJCNN)*, páginas 24–30, 2016. 29

[126] Lu, Ning, Jie Lu, Guangquan Zhang e Ramon Lopez de Mantaras: *A concept drift-tolerant case-base editing technique.* Artificial Intelligence, 230:108–133, 2016. 30

[127] Rad, Radin Hamidi e Maryam Amir Haeri: *Hybrid forest: A concept drift aware data stream mining algorithm.* Em *Cornell University: arXiv.* arXiv, 2019. 30

[128] Sethi, Tegjyot Singh e Mehmed Kantardzic: *On the reliable detection of concept drift from streaming unlabeled data.* Expert Systems with Applications, 82:77–99, 2017. 30

[129] Liu, Anjin, Jie Lu, Feng Liu e Guangquan Zhang: *Accumulating regional density dissimilarity for concept drift detection in data streams.* Pattern Recognition, 76:256–272, 2018. 30

[130] Xu, Shuliang, Lin Feng, Shenglan Liu e Hong Qiao: *Self-adaption neighborhood density clustering method for mixed data stream with concept drift.* Engineering Applications of Artificial Intelligence, 89:103451, 2020. 30

[131] Miyata, Yasushi e Hiroshi Ishikawa: *Concept drift detection on stream data for revising dbscan.* Electronics and Communications in Japan, 104(1):87–94, 2021. 30

[132] Berger, Vance W. e YanYan Zhou: *Kolmogorov–smirnov test: Overview.* Em *Encyclopedia of Statistics in Behavioral Science*, 2014. 30

[133] Wilcoxon, Frank: *Individual comparisons by ranking methods.* Journal of the American Statistical Association, 1(6):80–83, 1945. 30

[134] Vallim, Rosane M.M. e Rodrigo F. de Mello: *Proposal of a new stability concept to detect changes in unsupervised data streams.* Expert Systems with Applications, 41(16):7350–7360, 2014. 30

[135] Mouss, H., D. Mouss, N. Mouss e L. Sefouhi: *Test of page-hinckley, an approach for fault detection in an agro-alimentary production system.* Em *2004 5th Asian Control Conference (IEEE Cat. No.04EX904)*, páginas 815–818, 2004. 30

[136] Sun, Yongjiao, Ye Yuan e Guoren Wang: *An os-elm based distributed ensemble classification framework in p2p networks.* Neurocomputing, 74(16):2438–2443, 2011. 30

[137] Mahdi, Osama A., Eric Pardede, Nawfal Ali e Jinli Cao: *Diversity measure as a new drift detection method in data streaming.* Knowledge-Based Systems, 191:105–227, 2020. 30

[138] Krawczyk, Bartosz, Leandro L. Minku, João Gama, Jerzy Stefanowski e Michał Woźniak: *Ensemble learning for data stream analysis: A survey.* Information Fusion, 37:132–156, 2017. 30

[139] Wadewale, Kranti e Sharmishta Suhas Desai: *Survey on method of drift detection and classification for time varying data set.* International Research Journal of Engineering and Technology (IRJET, 2(9):709–713, 2015. 31

[140] Žliobaitė, Indrė, Albert Bifet, Jesse Read, Bernhard Pfahringer e Geoff Holmes: *Evaluation methods and decision theory for classification of streaming data with temporal dependence.* Machine Learning, 98(3):455–482, 2015. 32, 48

[141] Koychev, Ivan: *Experiments with two approaches for tracking drifting concepts.* Serdica Journal of Computing, 1(1):27–44, 2007. 32

[142] Klinkenberg, Ralf: *Learning drifting concepts: Example selection vs. example weighting.* Intelligent Data Analysis, 8(3):281—300, 2004. 32, 39, 49, 69

[143] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall e W. Philip Kegelmeyer: *Smote: Synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research, 16(1):321—357, 2002. 34, 35, 36, 37, 49, 50

[144] Galar, Mikel, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince e Francisco Herrera: *A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4):463–484, 2012. 34

[145] Han, Hui, Wen Yuan Wang e Bing Huan Mao: *Borderline-smote: A new over-sampling method in imbalanced data sets learning.* Em *Advances in Intelligent Computing*, páginas 878–887, 2005. 36, 37, 49

[146] Saito, Takaya e Marc Rehmsmeier: *The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets.* PLOS ONE, 10:1–21, 2015. 40, 43, 60, 68

[147] Japkowicz, Nathalie e Mohak Shah: *Evaluating Learning Algorithms: A Classification Perspective.* Cambridge University Press, 2011. 40

[148] Fawcett, Tom: *An introduction to roc analysis.* Pattern Recognition Letters, 27(8):861–874, 2006. 43

[149] Bradley, Andrew P.: *The use of the area under the roc curve in the evaluation of machine learning algorithms.* Pattern Recognition, 30(7):1145–1159, 1997. 43

[150] Hanley, J.A. e Barbara Mcneil: *The meaning and use of the area under a receiver operating characteristic (roc) curve.* Radiology, 143:29–36, 1982. 43

[151] He, Haibo, Yang Bai, Edwardo A. Garcia e Shutao Li: *Adasyn: Adaptive synthetic sampling approach for imbalanced learning.* Em *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, páginas 1322–1328, 2008. 49

[152] Nguyen, Hien M., Eric W. Cooper e Katsuari Kamei: *Borderline over-sampling for imbalanced data classification.* Int. J. Knowl. Eng. Soft Data Paradigm., 3(1):4—21, 2011. 49

[153] Batista, Gustavo E. A. P. A., Ronaldo C. Prati e Maria Carolina Monard: *A study of the behavior of several methods for balancing machine learning training data.* SIGKDD Explor. Newsl., 6(1):20—29, 2004. 49

[154] Batista, Gustavo E. A. P. A., Ana Lúcia Cetertich Bazzan e Maria Carolina Monard: *Balancing training data for automated annotation of keywords: a case study.* Em *WOB*, 2003. 49

[155] Chang, Chih Chung e Chih Jen Lin: *Libsvm: A library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2(3):1–27, 2011. 51

[156] Platt, John C.: *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.* Em *ADVANCES IN LARGE MARGIN CLASSIFIERS*, páginas 61–74. MIT Press, 1999. 51

[157] Breiman, Leo: *Random forests.* Machine Learning, 45(1):5–32, 2001. 51

[158] Breiman, Leo, Jerome H. Friedman, Richard A. Olshen e Charles J. Stone: *Classification And Regression Trees.* CRC Press, 1st edition edição, 1984. 51

[159] Chen, Tianqi e Carlos Guestrin: *Xgboost: A scalable tree boosting system.* Em *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 785—-794. Association for Computing Machinery, 2016. 51

[160] Carmona, Pedro, Francisco Climent e Alexandre Momparler: *Predicting failure in the u.s. banking sector: An extreme gradient boosting approach.* International Review of Economics & Finance, 61:304–323, 2019. 51

[161] Jabeur, Sami Ben, Cheima Gharib, Salma Mefteh-Wali e Wissal Ben Arfi: *Catboost model and artificial intelligence techniques for corporate failure prediction.* Technological Forecasting and Social Change, 166:120658, 2021. 51

[162] Hyndman, Rob J. e George Athanasopoulos: *Forecasting: Principles and Practice.* OTexts, 2021. 51, 63

[163] Demšar, Janez: *Statistical comparisons of classifiers over multiple data sets.* Journal of Machine Learning Research, 7:1—30, 2006. 62

[164] Friedman, Milton: *The use of ranks to avoid the assumption of normality implicit in the analysis of variance.* Journal of the american statistical association, 32(200):675–701, 1937. 62

[165] Friedman, Milton: *A comparison of alternative tests of significance for the problem of m rankings.* The Annals of Mathematical Statistics, 11(1):86–92, 1940. 62

[166] Nemenyi, Peter Bjorn: *Distribution-free multiple comparisons.* Princeton University, 1963. 62

Apêndice A

Indicadores econômicos e financeiros

A.1 Indicadores econômico-financeiros

A.1.1 Indicadores extraídos

Tabela A.1: Indicadores econômicos e financeiros.

Arquivo	**Indicador**	Conta origem
BPA	Ativo total	Ativo total
	Ativo circ.	Ativo circulante
	Disp.	Disponibilidades
	Receb.	Recebíveis
	Estoque	Estoque
	Ativo não-circ.	Ativo não-circulante
	Ativos intang.	Ativos intangível
	Ativo permanente	Ativo permanente
	Deprec. acumulada	Depreciação acumulada
	Amortiz. acumulada	Amortização acumulada
	Investimentos	Investimentos
BPP	Passivo total	Passivo total
	Passivo circ.	Passivo circulante
	Passivo não-circ.	Passivo não-circulante
	Exigível	Passívo circ. − Passivo não-circ.
	Patrimônio líq.	Pass total − (Pass circ. + Pass não-circ.)
	Capital social	Capital social Realizado
	Reservas	R de lucros + R de capital + R de reavaliação
	Provisões	Provisões

	Empres. de longo prz	Empréstimo de longo prazo
DRE	Receita bruta	Receita Bruta
	Custos	Custos
	Receita líquida	Receita líquida (LAJIRDA)
	Despesas oper.	Despesas operacionais
	Resultado oper.	Resultado operacional (LAJIR)
	Resultado finan.	Resultado financeiro
	Despesas finan.	Despesas financeira
	Lucro após result finan.	Lucro após resultado financeiro (LAIR)
	Desp. com imposto	Despesas com imposto
	Resultado líq.	Resultado líquido
DFC	Caixa oper.	Caixa operacional
	Caixa invest.	Caixa de Investimento
	Caixa finan.	Caixa Financeiro

A.1.2 Indicadores calculados

A.1.3 Liquidez de curto prazo

Liquidez corrente

$$\frac{Ativo\ circulante}{Passivo\ circulante}$$

Liquidez seca

$$\frac{(Ativo\ circulante\ -\ Estoque)}{Passivo\ circulante}$$

Liquidez imediata

$$\frac{(Caixa\ +\ Caixa\ Equivalente)}{Passivo\ circulante}$$

A.1.4 Liquidez de longo prazo

Índice de cobertura de juros

$$\frac{EBIT}{Despesas\ com\ juros}$$

Índice de endividamento total

$$\frac{Passivo\ circulante + Passivo\ não\text{-}circulante}{Ativo\ total}$$

Índice de cobertura de ativos tangíveis

$$\frac{Ativo\ total - Ativo\ intangível - Passivo\ circulante}{Ativo\ total}$$

Relação patrimônio e dívida

$$\frac{Ativo\ total - (Passivo\ circulante + Passivo\ não\text{-}circulante)}{Passivo\ circulante + Passivo\ não\text{-}circulante}$$

Dívida total pelos ativos totais

$$\frac{Exigível\ total}{Ativos\ tangíveis}$$

A.1.5 Estrutura dos ativos

Taxa de liquidez

$$\frac{Ativo\ circulante + Realizável\ a\ longo\ prazo}{Passivo\ circulate + Passivo\ não\text{-}circulante}$$

Índice de ativos a receber

$$\frac{Contas\ a\ receber}{Ativo\ circulante}$$

Relação de ativos fixos

$$\frac{(Ativos\ tangíveis - Deprecisação\ acumulada)}{Cap.\ social + Reservas + Empréstimo\ de\ longo\ prazo}$$

Índice de patrimônio líquido para ativos fixos

$$\frac{Patrimônio\ líquido\ líquido}{Ativos\ fixos\ totais}$$

Liquidez corrente

$$\frac{Passivo\ circulante}{Ativo\ circulante}$$

A.1.6 Capacidade operacional

Margem de lucro

$$\frac{Lucro\ bruto}{Receita\ total}$$

Relação recebíveis e receita

$$\frac{Recebíveis}{Receita\ total}$$

Relação estoque e lucro

$$\frac{Estoque}{Resultado\ bruto}$$

Giro de estoque

$$\frac{Custo\ de\ mercadorias\ vendidas}{Valor\ médio\ de\ estoque}$$

Giro de contas a pagar

$$\frac{Valor\ total\ das\ compras}{Valor\ médio\ das\ contas\ a\ pagar}$$

Giro do ativo circulante

$$\frac{Receita\ bruta}{Valor\ médio\ do\ ativo\ circulante}$$

Relação ativo permanente e lucro

$$\frac{Tota\ de\ ativos\ fixos}{Receita\ líquida}$$

Giro do capital total

$$\frac{Receitas}{Patrimônio\ líquido}$$

A.1.7 Lucratividade

Retorno sobre ativos (ROA)

$$\frac{Resultado\ líquido}{Ativo\ total\ médio}$$

Relação lucro líquido e ativos totais

$$\frac{Lucro\ líquido}{Ativo\ total\ médio}$$

Relação resultado líquido e ativos circulante

$$\frac{Resultado\ líquido}{Ativo\ circulante}$$

Relação lucro líquido e ativos permanente

$$\frac{Lucro\ líquido}{Ativos\ permanente}$$

Retorno sobre patrimônio (ROE)

$$\frac{Resultado\ líquido}{Patrimônio\ líquido}$$

Margem operacional

$$\frac{Lucro\ operacional}{Receita\ líquida} \times 100$$

Relação custo oper. total e receita bruta

$$\frac{Custo\ operacionais}{Receita\ bruta}$$

Relação de despesas e receitas

$$\frac{Despesas\ totais}{Receitas\ totais} \times 100$$

Índice de despesas de gestão

$$\frac{Despesas\ administrativas}{Receitas\ totais} \times 100$$

Índice de endividamento

$$\frac{Despesas\ financeiras}{Receitas\ totais} \times 100$$

A.1.8 Fluxo de Caixa

Fluxo de Caixa Livre

$$Fluxo\ de\ caixa\ oper. - Despesa\ de\ Capital$$

Relação fluxo de caixa operacional e lucro líquido

$$\frac{Caixa\ operacional}{Lucro\ líquido}$$

Relação de fluxo de caixa operacional e receita

$$\frac{Caixa\ operacional}{Receita\ bruta}$$

Taxa de recuperação de caixa

$$\frac{Caixa\ operacional}{Ativo\ total\ médio} \times 100$$

A.1.9 Nível de risco

Grau de alavancagem financeira

$$\frac{Retorno\ sobre\ patrimônio\ líquido}{Retorno\ sobre\ ativo}$$

Grau de alavancagem operacional

$$\frac{\Delta\%\ LAJIR}{\Delta\%\ da\ Receita} = n^o\ de\ vendas$$

Grau de alavancagem operacional

$$\frac{\Delta\% Lucro\ líquido}{\Delta\% de\ Receita} = n^o\ de\ vezes$$

A.1.10 Capacidade de crescimento

Taxa de crescimento de manutenção de capital

$$\frac{MC_t\ -\ MC_{t-1}}{MC_{t-1}}$$

Taxa de crescimento de capital acumulado (CA)

$$\frac{CA_t\ -\ CA_{t-1}}{CA_{t-1}}$$

Taxa de crescimento de ativos totais

$$\frac{Ativos\ totais_t\ -\ Ativos\ totais_{t-1}}{Ativos\ totais_{t-1}}$$

Taxa de crescimento do ROE

$$\frac{ROE_t\ -\ ROE_{t-1}}{ROE_{t-1}}$$

Taxa de crescimento do lucro líquido

$$\frac{Lucro\ líquido_t - Lucro\ líquido_{t-1}}{Lucro\ líquido_{t-1}}$$

Taxa de crescimento do lucro operacional

$$\frac{Lucro\ operacional_t\ -\ Lucro\ operacional_{t-1}}{Lucro\ operacional_{t-1}}$$

Taxa de crescimento das receitas operacionais

$$\frac{Receita\ operacional_t\ -\ Receita\ operacional_{t-1}}{Receita\ operacional_{t-1}}$$

Taxa de crescimento do custo (de operação)

$$\frac{Custo\ operacional_t\ -\ Custo\ operacional_{t-1}}{Custo\ operacional_{t-1}}$$

A.1.11 Indicador por ação

Lucro por ação

$$\frac{Lucro\ líquido}{Ações\ em\ circulação}$$

Valor de ativos líquidos por ação

$$\frac{Ativo\ circulante - (Passivo\ circulante + Passivo\ não\text{-}circulante)}{Ações\ em\ circulação}$$

Caixa livre por ação

$$\frac{Caixa\ livre}{Ações\ em\ circulação}$$