Enhancing Financial Distress Prediction with Preprocessing and Ensemble Methods: A Comparative Study

Mithul Murugaadev
Student ID: 1130073

Thesis Report for
Master of Science in Data science

Liverpool John Moores University

September 2024

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Nilam Upasani for all her help and advice with this thesis. I would also like to thank my parents for their support. I also extend my gratitude to my fellow students in my cohort for their support. Lastly, I would like to express my thanks to LJMU and UpGrad for providing the studentship that enabled me to carry out this thesis.

ABSTRACT

The ability to accurately predict financial distress is crucial for various stakeholders, including investors, creditors, and regulatory bodies, as it enables timely interventions and informed decision-making. However, the efficacy of predictive models is heavily influenced by the quality of the underlying data. This thesis delves into the critical challenges of data quality, specifically focusing on null value imputation and class imbalance, and examines their impact on the performance of financial distress prediction models.

The research begins by addressing the pervasive issue of missing data, a common occurrence in financial datasets that can significantly distort predictive outcomes. Advanced imputation methods are employed to reconstruct incomplete datasets, aiming to preserve the integrity and informational value of the data. Following this, the thesis investigates the challenge of class imbalance, where the disproportionate representation of distressed versus non-distressed firms can lead to biased predictions. To counteract this, the study applies sophisticated resampling techniques, chosen for their unique ability to handle complex data structures and improve model robustness.

By integrating these preprocessed datasets with various machine learning algorithms, the research provides a comprehensive analysis of the influence of data quality on prediction outcomes. The findings highlight the critical role of data quality management in enhancing model performance.

This thesis makes a significant contribution to the field by providing insights into the relationship between data quality challenges and the effectiveness of predictive models. The findings offer a solid foundation for future research and practical applications in financial risk assessment, highlighting the importance of addressing data quality issues to enhance predictive accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The study of financial distress, characterized by a company's inability to meet financial obligations due to factors like declining revenues, profitability, and liquidity, has long intrigued researchers (Ramzan, 2023). Key predictive methods include bankruptcy prediction and credit scoring, which initially relied on expert judgment. Over time, statistical models like those by Altman, Ohlsen, and Beneish, as well as techniques such as logit, probit, and linear probability, have become popular for predicting corporate collapse. However, these models face limitations related to assumptions, multicollinearity, outliers, and missing data, prompting the adoption of machine learning algorithms (Gabrielli et al., 2023).

Recent advances have led to more accurate prediction models, including hybrid models combining PCA with ANN, Sparse PCA with SVM, and CNN-based models using company reports instead of just quantitative data (Adisa et al., 2019; Matin et al., 2019; Zeng and Yang, 2020). However, issues like missing data and class imbalance continue to challenge the reliability of financial distress prediction models. Distressed firms often represent a minority, leading to biased models that favor the majority class, impacting accuracy (Garcia, 2022; Hassan and Yousaf, 2022).

This thesis focuses on enhancing financial distress prediction by addressing data quality issues, such as missing data and class imbalance, using null value imputation and advanced techniques. Improving predictive accuracy benefits stakeholders, including investors, entrepreneurs, policymakers, and governmental organizations, by supporting informed decision-making and planning. Ultimately, effective financial distress prediction safeguards economic stability and financial security for individuals and businesses alike.

1.2 Research Questions

This research tries to answer the following questions:

1. What machine learning techniques are most effective in the imputation of missing values in high-dimensional accounting and financial data?
2. What are the most effective class imbalance techniques that are suitable for preprocessing the financial data?
3. What are the comparative performances of various ensemble tree methods and other advanced machine learning algorithms in financial distress prediction?

1.3 Aim & Objectives

The research is conducted to develop a robust and comprehensive machine learning framework for predicting financial distress in companies using high-dimensional accounting and financial data, incorporating various techniques to address missing values, class imbalance, dimensionality reduction, and ensemble modeling.

The research objectives are formulated based on the aim of this study, which are as follows:

- To preprocess high-dimensional data exploring suitable techniques for noisy data and missing values.
- To Investigate and implement appropriate class imbalance techniques.
- To build predictive models to identify the most accurate and high-performing model to classify distressed companies based on financial and accounting data.
- To evaluate the performances of the classifiers using comprehensive evaluation metrics

and techniques.

1.4 Significance of the Study

This study holds significant implications for academia, industry, and regulatory bodies alike. By employing machine learning techniques and frameworks to predict financial distress, it aims to enhance financial risk management practices and ensure the sustainability of businesses.

Early detection and proactive mitigation of financial distress are crucial for safeguarding investments and maintaining business continuity. Through the development of robust predictive models with suitable preprocessing frameworks, this study empowers decision-makers to make informed strategic decisions, optimize resource allocation, and navigate financial challenges effectively.

The interdisciplinary nature of this research fosters collaboration and innovation across finance, data science, and regulatory domains. By facilitating knowledge exchange and continuous improvement, it drives progress and adaptation in an evolving global landscape.

Lastly, this study expands theoretical understanding and empirical evidence in financial risk management.

1.5 Scope of study

The scope of this thesis work is defined as follows:

- The thesis will focus on the development and evaluation of machine learning models and methodologies for financial distress prediction using high-dimensional accounting and financial data.
- The data utilized in this research are derived from accounting and financial statements, including balance sheets, income statements, and cash flow statements.

1.6 Structure of the study

This thesis is organized into six chapters, each designed to systematically explore and address the research objectives.

- Chapter 1 – introduction: This chapter introduces the research topic, highlights the significance of financial distress prediction, addresses the challenges of data quality, and outlines the research objectives. It sets the stage for the thesis by presenting the research questions.

- Chapter 2 – Literature review: This chapter reviews existing literature on financial distress prediction and business failure prediction, highlighting key works in the field. It discusses traditional and modern approaches, and also addresses previous research on the challenges of financial data, situating the current study within the broader academic context.

- Chapter 3 – Methodology: The methodology chapter outlines the research design and analytical framework employed in this study. It provides a detailed explanation of the chosen techniques for the study. It articulates the rationale behind these methods and describes the machine learning models and evaluation metrics used to assess predictive performance.

- Chapter 4 – Analysis: This chapter presents the exploratory data analysis (EDA) and the technical execution of the research methodology. It details the process of implementation of the techniques, and training and validation of predictive models.

- Chapter 5 – Results and Discussion: In this chapter, the results of the empirical analyses are systematically presented and interpreted. The chapter evaluates the impact of the data preprocessing strategies on model performance. It discusses the implications of the results, providing an understanding of the effectiveness of the methods employed and their relevance to financial distress prediction.

- Chapter 6 – Conclusion: The concluding chapter synthesizes the key findings of the thesis, and concludes the thesis. It also identifies potential avenues for future research.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter reviews the existing literature in the areas of financial distress prediction and bankruptcy prediction.

2.2 Analytics in Finance

There has been ongoing and significant advancement in the application of analytics and machine learning across various sectors of the finance domain. Researchers and practitioners are continuously exploring sophisticated models to enhance predictive capabilities in these fields.

2.2.1 Stock Market Prediction

Through their study, (Liu et al., 2021) investigate the impact of investors' social interactions on stock prices by incorporating social network variables from social media into predictive models. They utilize Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Recurrent Neural Network (RNN) models. The results demonstrate that these social network variables significantly improve prediction accuracy across all models, with LSTM slightly outperforming GRU and RNN. This study underscores the importance of investor social interactions, as captured through shared attention on social media, for enhancing stock price forecasts. While, (Pokhrel et al., 2022) in their study, develop models using sixteen predictors, including fundamental, macroeconomic, technical, and financial news data, to predict stock prices. They compare LSTM, GRU, and CNN architectures for forecasting the NEPSE index, optimizing hyperparameters. Results show the LSTM model with 30 neurons outperforms GRU and CNN, based on RMSE, MAPE, and R metrics. The LSTM model's effectiveness is validated through statistical analysis.

LSTM has proven effective for short-term predictions, as demonstrated by (Banik et al., 2022) Decision Support System (DSS) for Indian swing traders. This system integrates the Investment Success Score, technical indicators, and LSTM predictions to enhance stock market forecasting.

LSTM outperforms other models, with MACD-Signal Line analysis, MFI, and RSI offering valuable buy or sell signals, while Support-Resistance curves and Fibonacci retracement levels provide insights for various trading durations. Despite its effectiveness, the DSS's high computational demands may limit its use for scalp trading on less powerful systems, though this can be mitigated with high-performance computers with dedicated GPUs.

2.2.2 Cryptocurrency

The rise of bitcoin trading and usage has also increased the demand of advanced analytics and heightened security. This growth has led to the development of sophisticated trading algorithms, enhanced security protocols, and the need for more robust regulatory frameworks to manage risks associated with the decentralized and volatile nature of cryptocurrencies. Hence, various studies are exploring the prediction and advanced analytics scope of cryptocurrency due to their high volatility.

Regarding this context, (Rathore et al., 2022) in their research, address Bitcoin value forecasting amid cryptocurrency market volatility. The study compares the fbprophet model with a Naive model, highlighting fbprophet's superior performance in handling seasonal trends, missing data, and outliers. Results show that fbprophet's ability to detect and adjust for seasonal patterns makes it a robust tool for real-world cryptocurrency forecasting. In a contrasting trajectory, (Wang et al., 2022b) examines how investor trading behaviors, particularly informed trading, impact return predictability in the cryptocurrency market. Analyzing 12 major cryptocurrencies, they find that while informed trading indicators can predict returns for some cryptocurrencies, they do not significantly improve overall market predictability. Other investor behavior indicators also fail to enhance prediction accuracy. The study suggests that the cryptocurrency market is chaotic, influenced by uninformed traders, macroeconomic factors, and external events.

Alternatively, Another study on Ethereum price prediction emphasizes the importance of commodity-specific variables. (Kim et al., 2021) finds that macroeconomic factors and Ethereum-specific Blockchain variables, such as uncle blocks, gas prices, gas consumption, and gas limits, significantly improve prediction accuracy. By incorporating these factors along with general Ethereum Blockchain data and Bitcoin's Blockchain information, the study achieves the best predictive results. Artificial Neural Networks (ANN) are noted for their strong performance in predicting Ethereum prices.

2.2.3 Portfolio Management

Advancements in Machine learning has also helped in optimization of asset allocation, mitigating behavioral bias in investment decisions by continuously analyzing real-time data and adjusting asset weights to optimize returns while controlling for risk. Diversification and strategic planning are the areas that demand high focus in portfolio management. Studies prove that ML models are robust in generating portfolios that outperform the local market. In relative terms, this research focuses on developing portfolios that specifically address and manage the high volatility associated with short positions.

To address this, (Rubesam, 2022)introduces an Equal Risk Contribution (ERC) approach to balance risk between long and short positions, enhancing risk-adjusted returns using technical and fundamental indicators. The study develops a multi-strategy ERC approach that combines ML-based long-short strategies, equalizing risk across strategies. This approach outperforms individual ML strategies, ensemble forecasts, and equal-weighted multi-strategy methods. (Pinelis and Ruppert, 2022) explores the effectiveness of Random Forest and Elastic Net ML models in portfolio allocation, showing significant improvements in risk-adjusted returns and utility gains. The study examines return- and volatility-timing strategies, highlighting how ML enhances portfolio performance and alpha generation in actively managed portfolios. The findings underscore ML's substantial benefits in finance, even with standard variables, beyond big data contexts.

(Ma et al., 2021) extends portfolio construction literature by incorporating ML and DL models into MV and omega portfolio optimization frameworks, focusing on China Securities 100 Index stocks from January 2012 to December 2015. The study compares Random Forest (RF), Support Vector Regression (SVR), Deep Multi-Layer Perceptron (DMLP), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN) for stock return predictions. Results show that RF outperforms other models, with RF MVF and RF OF being the most effective. Despite high turnover, RF MVF remains the top performer even after accounting for transaction fees. The study recommends RF MVF for daily trading investments.

2.2.4 Forex

Foreign exchange (forex) trading involves the exchange of currencies in a global marketplace. The traders seek to profit from fluctuations in currency exchanges. ML applications in forex have significantly advanced the analysis and trading strategies in this market. Due to its high liquidity in currency exchange rates, Forex trading presents both significant opportunities and challenges.

In this study, (Peng and Lee, 2021) developed a log-distance path loss model to measure data noise impacts and address overfitting in algorithmic trading. This model helped select optimal trading exchange pairs and frequencies to maximize profit. Results showed that holding a position until a switch was required outperformed single-period holding, validating FX trading as a Markov Decision Process (MDP). Experiments with path loss metrics, regression objectives, and fine-tuned hyperparameters demonstrated positive out-of-sample returns, confirming the effectiveness of the proposed approach across various ML methods. Widening the scope, (Hassanniakalager et al., 2021) explores technical analysis and Bayesian statistics in trading, guided by the Adaptive Market Hypothesis (AMH). They generate 7,846 technical trading rules for EUR/USD, GBP/USD, and USD/JPY and identify 5% to 15% as genuinely profitable using data snooping and balancing procedures. These rules yield modest returns and Sharpe ratios. When combined with Bayesian models—Naive Bayes, Relevance Vector Machine, Dynamic Model Averaging, Dynamic Model Selection, and Bayesian Neural Networks—the trading

performance improves significantly, achieving 6% annual returns after transaction costs. DMA and BNN notably outperform benchmarks, highlighting the effectiveness of Bayesian techniques.

(Ahmed et al., 2020) proposes a system integrating the Forex Loss Function (FLF) with Long Short-Term Memory (LSTM) networks for EUR/USD Forex prediction. This approach improves prediction accuracy by 10.96% over classic LSTM models, reducing mean absolute error for the next H4 candle's opening price by 19.19% and for closing, high, and low prices by 10.54%, 7.05%, and 7.08%, respectively. FLF-LSTM outperforms FLF-RNN, cutting forecasting error by 73.57%, and shows 13% and 37.7% error reductions compared to ARIMA and FB Prophet, respectively. The study underscores the value of domain-specific knowledge in enhancing LSTM performance and the usage efficiency and accuracy of LSTM as seen with previous studies.

2.2.5 Financial Crisis

Early detection of financial crisis have been highly focused subject. Models are employed to identify early warning signs and understand the complex factors that contribute to financial instability.

(Tölö, 2020) investigated systemic financial crisis prediction using neural networks and the Jordá-Schularick-Taylor dataset (1870–2016), finding that time series inputs enhance prediction accuracy. Recurrent neural networks (RNNs), particularly RNN-LSTM and RNN-GRU, outperformed traditional logit models and multilayer perceptron (MLP) models, showing superior out-of-sample performance for forecasts up to five years. LSTM models provided reliable signals at forecast horizons. Key predictors, including stock prices, loans/GDP, house prices, current account/GDP, and GDP, were crucial, with stock prices being especially influential.

On a varied approach relying on textual data, (Petropoulos and Siakoulis, 2021) explores the predictive power of central bank speeches on medium-term financial turmoil using natural language processing (NLP). The study innovates with an automatically adjustable dictionary to

enhance framework flexibility and accuracy. The sentiment index derived from this analysis acts as an early warning tool for global financial instability. Results show XGBoost outperforms other ML techniques, such as Deep Neural Networks, Support Vector Machines, and random forests, in AUROC and KS metrics, highlighting the value of diverse approaches in improving forecasting accuracy.

The mentioned research highlights the various areas of applications of machine learning in finance. It also has been observed that LSTM is a highly efficient model as stated by various research papers suitable for a wide range of advanced prediction analytics.

2.3 Data Mining in Financial Risk Assessment

Financial institutions face complex and dynamic environments where risks can arise from various sources, including credit default, market volatility, fraud, and operational failures. The sheer volume and complexity of financial data make it difficult to detect emerging risks and anomalies using conventional techniques. Data mining helps solve these problems by enabling the detection of subtle patterns and correlations within large datasets that might otherwise go unnoticed. Data mining is instrumental in fraud detection by analyzing transaction patterns to flag unusual activities that could indicate fraudulent behavior.

(Huang et al., 2021) focuses on predicting enterprise risk for loan applicants using financial ratios. They compare Random Forest (RF), Support Vector Machine (SVM), and AdaBoost, finding AdaBoost to be the highest-performing model with 90.1% accuracy and superior Area Under the Curve (AUC) compared to RF. SVM performs the weakest in accuracy and AUC. Noise robustness tests reveal that RF and AdaBoost are more resilient to data noise than SVM. Focusing on the macro level, (Wang et al., 2024a) investigates systemic risk in the Chinese financial sector by analyzing network connectedness among publicly-listed financial institutions. Using the CoVaR (Conditional Value at Risk) approach, the study builds dynamic correlation networks to capture institutional interconnectedness. Four machine learning models are

employed to predict systemic risk across various time horizons, with and without network connectedness as input. The study also uses a fingerprint model to analyze systemic risk drivers, including linear, nonlinear, and interaction effects.

While assessing risk is crucial for any organization, (Fan et al., 2023) in their study, develop a financial risk assessment model for football clubs using data from 21 clubs during the 2018-2019 season. The study employs financial ratios, Principle Component Analysis (PCA) for feature extraction, and Exploratory Factor Analysis (EFA) for model construction. Findings reveal significant financial risks, including small capital size, high asset-liability ratios, low net profit, substantial losses, and weak asset liquidity. The research highlights the critical role of financial ratios in assessing financial risk. (Murugan and T, 2023) introduces a novel credit risk assessment model integrating Internet of Things (IoT) technology to analyze client data. They evaluate K-Nearest Neighbors , Logistic Regression , and XGBoost classifiers, finding KNN effective despite challenges in selecting the optimal number of neighbors. The model uses information gain analysis to identify key predictive attributes and shows superior performance compared to existing methods. Simulation results indicate that the model achieves a wealth proportion measure of 0.02 to 0.09 and employs a value-at-risk strategy to maintain optimal consumption stability below 5% of total investment wealth.

(Duan, 2019) targets risk prediction in peer-to-peer (P2P) lending using data from 2007 to 2015, addressing class imbalance with Synthetic Minority Over-sampling Technique (SMOTE). They employ a Deep Neural Network (DNN) with a Multi-Layer Perceptron (MLP) architecture, featuring three hidden layers and 28 borrower-related features, to predict loan defaults. The MLP model achieves a high prediction accuracy of 93.18%, outperforming traditional logistic regression and simpler MLP models. Sensitivity and specificity values of 75.6% and 72.2%, respectively, demonstrate robust performance in identifying both default and non-default cases. The study underscores the effectiveness of combining SMOTE and a well-structured MLP for reliable P2P lending risk prediction.

These studies emphasize the importance of conducting thorough risk assessments and showcase the robust methodologies that can be employed in this process. When these risks materialize, they often lead to financial distress.

2.4 Introduction to Financial Distress Prediction

Financial risk assessment is crucial for evaluating threats to an organization's stability, with poor management often leading to financial distress when companies struggle to meet financial obligations. Traditionally, predicting financial distress relied on ratio and financial statement analysis using historical data. However, the rise of machine learning has revolutionized this field by employing algorithms to detect patterns traditional methods may overlook. Advanced techniques like deep learning and ensemble models have further improved prediction accuracy and robustness, reflecting increased sophistication due to greater data availability and computational power.

2.4.1 Traditional Methods

(Kottala and Sahu, 2024) through their research, investigate the intersection of ergonomics and financial distress, specifically within the manufacturing sector, to enhance organizational decision-making and efficiency. Their research highlights how understanding employees' fitness and working conditions can significantly impact financial distress management. By integrating ergonomic assessments with financial distress analysis, the study suggests that organizations can develop better policies and regulations that not only improve workplace conditions but also enhance overall financial performance.

Delving further into the methods, (Dinh et al., 2021) evaluates prediction among companies in ASEAN countries using three key indicators: Interest Coverage Ratio (ICR), Non-Performing Loans (NPL), and Distance to Default (DD). Leveraging data from Bloomberg and Datastream, the study focuses on market-based models over traditional accounting-based ones. The DD model proves to be the most robust, accurately forecasting financial distress up to one year in

advance and outperforming ICR and NPL in many cases. The research highlights the effectiveness of market-based models as a valuable alternative to conventional accounting-based approaches.

Further narrowing the context to liquidity, a subset of indicators is particularly well-suited for assessing distress in banks because they specifically address key aspects of financial stability and solvency, such as cash flow, asset liquidity, and short-term financial obligations, covering core operations of banks, (Chen et al., 2022) in their study evaluate the effectiveness of three liquidity indicators—Liquidity Ratio (LiqR), Liquidity Creation (LiqC), and Net Stable Funding Difference (NSFD)—in predicting early signs of distress among banks in the United States and the European Union. The research finds that LiqC and NSFD are superior to LiqR as early warning signals. LiqC and NSFD consistently provide more reliable and stable reflections of bank liquidity performance and potential weaknesses, while LiqR's signals are less dependable and sensitive to varying specifications.

(Leng and Sun, 2024) utilizes Altman's Z score to predict financial distress among Chinese A-share listed companies, focusing on the impact of COVID-19. The research finds that the pandemic has worsened conditions in sectors such as mining, manufacturing, and transportation, especially for small-scale enterprises in central and western regions. In contrast, sectors like information transmission, software, and IT services have benefited. The study highlights that digital transformation and government subsidies are crucial in mitigating the pandemic's adverse effects. The Z score proves effective in quantifying distress and assessing how external shocks like COVID-19 influence financial health, demonstrating the value of integrating financial metrics with an understanding of external impacts for enhanced distress prediction.

Moreover, Authors propose more refined methods for specific use cases. In that regard (Çolak, 2021) introduced the Multivariate Firm Assessment (MFA) score to improve FDP for nonfinancial firms listed on the Borsa Istanbul (BIST) from 2001 to 2017. The MFA score integrates multiple financial ratios and outperforms traditional methods like Multi-Discriminant

Analysis and the Altman Z-score, achieving an average accuracy of 92%. It effectively reflects macroeconomic influences, correlates with GDP growth, exchange rates, and industrial production, and serves as an early warning system The study reveals that firms with open foreign exchange positions have lower MFA scores, indicating higher vulnerability, and smaller firms generally face more distress than larger firms, which manage foreign exchange risks better.

Comparably, (Lohmann and Möllenhoff, 2023) presents a Bankruptcy Risk Matrix to improve bankruptcy prediction for US-listed companies. This matrix comprises two components of which, one measures bankruptcy risk and the other tracks recent changes in this risk. It visualizes and interprets outcomes from various prediction models by mapping risk on a scale from 0 to 1, applicable to an entire market or specific subsamples. Along a similar trajectory, adding economic components with financial variables, (Figlioli and Lima, 2022) explores the FL-score for predicting FDP using data from U.S. non-financial public firms from 2002 to 2019. The FL-score, combining financial and economic components, effectively classifies types of financial distress and proves robust across sectors. The study shows that integrating the FL-score with machine learning models like logistic regression and LASSO enhances both generalization and accuracy.

These studies cover a range of methods for distress prediction, but traditional models struggle with large datasets, fail to capture unique or nonlinear patterns, and lack adaptability to new conditions. In contrast, Machine Learning (ML) models excel in handling large datasets, identifying complex patterns, adapting to varying conditions, and learning from new data, making them ideal for advanced predictive tasks.

2.4.2 Machine learning in Financial Distress Prediction

Machine learning models have emerged as powerful tools in FDP, offering enhanced predictive capabilities over traditional statistical methods. Models like logistic regression are highly used in distress prediction in combination with traditional metrics. In a related vein, (Liu et al., 2023)

explores predicting financial distress by incorporating jump-tail risk into a logistic regression model, using data from Chinese stock exchanges.

Delving deeper in exploration, this study conducted by (Tang et al., 2024), aims to predict financial distress in the market using data from multiple sub-markets obtained from the WIND database, including stock, bond, credit, money, and foreign exchange markets. Employing a Markov regime-switching model, the study identifies distress periods effectively, with submarket stress indices proving more accurate than traditional models. The research also compares forecasting performance between traditional models (GLM, Logistic Regression) and advanced models (XGBoost, Decision Trees), finding that advanced models excel with complex datasets and feature variables. In the context of comparisons, (Putri and Dhini, 2019) compares conventional statistical methods, like logistic regression, with data mining techniques, including decision trees, support vector machines (SVM), and ensemble methods such as bagging and boosting. The research highlights that decision tree models, particularly when enhanced with boosting techniques, achieved the highest prediction accuracy of 94.61%, alongside superior sensitivity and specificity (94%), and minimized type II errors (5.5%).

Further exploring the efficiency of models, (Barboza and Altman, 2024) investigates financial distress prediction across Latin America, comparing Logistic Regression and Random Forest models using data from Argentina, Brazil, Chile, Colombia, Mexico, and Peru. The study examines how the relevance of financial indicators has shifted before and after the COVID-19 pandemic. It finds that while indicator importance has remained consistent over the past 20 years, LR and RF models show varying performance based on these indicators. Investigating ensemble methods, in a study performed by (Rahayu and Suhartanto, n.d.), specifically Random Forest and AdaBoost, were evaluated for predicting financial distress using data from Indonesian public companies. The study compared these ensemble techniques with single machine learning methods, using six financial variables and testing two bankruptcy class groupings. Results showed AdaBoost outperformed Random Forest in both accuracy and stability.

Building on the ensembles, (Rahman and Zhu, 2024) explores explores the capabilities of CUSBoost in FDP among China-A listed construction companies. By comparing CUSBoost with traditional Z-score models, the study highlights CUSBoost's superior handling of class imbalance issues in the dataset. Using financial ratios as input variables and applying Principal Component Analysis (PCA) to manage correlated variables, CUSBoost achieved high average AUC scores, demonstrating its robustness and enhanced predictive accuracy. In a similar trajectory, (Jabeur et al., 2021) assessed the CatBoost algorithm for FDP among French firms using data from 2014 to 2016. Compared to eight other algorithms, CatBoost excelled in prediction accuracy, particularly with categorical variables.

In extension of that regard, (Lokanan and Sharma, 2024) improved financial statement fraud detection by integrating machine learning algorithms with criminological theories. They utilized Recursive Feature Elimination with cross-validation (RFECV) for feature selection, Principal Component Analysis (PCA) for feature extraction, and Classification and Regression Tree (CART) with bootstrapping for prediction.

Following the approach of integrating additional variables and exploring a range of predictive models, (Citterio and King, 2023) examines the role of Environmental, Social, and Governance (ESG) scores in enhancing FDP for banks, alongside traditional financial ratios and macroeconomic variables like GDP, inflation (INF), and the Herfindahl-Hirschman Index (HHI). Analyzing data from US and EU-28 listed banks (2012-2019), their research showed that ESG scores significantly improve FDP model accuracy, boosting the Area Under the Curve (AUC) by up to 2.1 points and the F-score by up to 4.1 points. Notably, ESG integration reduced Type II errors, decreasing misclassification of distressed banks as healthy and mitigating costly false negatives. Various ML models incorporating ESG variables were also explored.

Expanding the comparison across a broad range of ML models (Carmona et al., 2019) investigated bank failure prediction in the U.S. banking sector (2001-2015) using XGBoost. The study found XGBoost significantly outperforms traditional models like Logistic Regression and Random

Forest in predictive accuracy. By employing a case-control matching strategy to balance data between failed and non-failed banks, XGBoost's superior performance is attributed to its handling of complex data interactions. Supplementing the approaches, (Son et al., 2019) compared various ML models and employed key preprocessing techniques to enhance data quality. They used zero imputation for missing values, winsorization to manage outliers, and Box-Cox transformations to correct skewness. The research found that XGBoost and Artificial Neural Networks (ANN) were the most effective. XGBoost excelled with its gradient boosting capabilities, handling complex interactions and non-linearities, while ANN effectively captured intricate patterns and dependencies.

On a parallel course, (Aydin et al., 2022) delves into the predictive capabilities of Artificial Neural Networks (ANN) and Decision Trees (DT) by incorporating both traditional financial metrics and non-financial variables, such as the number of employees. Enhancing the depth of the research on Neural networks, (Brenes et al., 2022) investigates Multi-Layer Perceptron (MLP) models for bankruptcy prediction using a Taiwanese dataset of 95 financial ratios from 1999 to 2009. The study evaluates various MLP configurations, finding the Adam optimization algorithm superior to SGD, though no activation function had a clear edge. MLP models achieved 83% accuracy, with models M2 and M22 showing the highest specificity; M22 was preferred for its simplicity. A balanced subsample of 550 firms was selected from 6,599 non-bankrupt firms to address class imbalance. MLP models outperformed top Support Vector Machine (SVM) models in similar studies.

2.4.3 Hybrid Strategies for Distress Prediction

ML models have proven to be highly effective in predicting financial distress, demonstrating their strengths in overcoming the limitations of traditional methods. Additionally, research indicates that integrating various techniques has further enhanced accuracy. Many studies have explored this domain by combining different methods and an ensemble of models to optimize performance.

Exploring this context, in their study, (Yi et al., 2023) investigate predicting financial risk in supply chain finance, particularly focusing on liquidity crises faced by suppliers due to delayed payments. The research employs various machine learning algorithms, including Random Forest (RF), XGBoost, Gradient Boosting Decision Tree (GBDT), and LightGBM, to model financial risk. Among the models tested, XGBoost stands out as the most effective, demonstrating superior performance in predicting financial risk related to supply chain finance. Expanding on the XGBoost model, in their study, (Liu et al., 2022) utilizes an advanced XGBoost model with data from the China Security Market Accounting Research Database (CSMARD). The research introduces a weighted cost-sensitive XGBoost approach, specifically developing a weighted XGBoost-based tree (XGBoost-W-BT) to address class imbalance issues. This model incorporates weighted costs to better handle the unequal distribution between distressed and non-distressed instances, aiming to reduce misclassification errors and improve prediction accuracy.

Broadening further on the model, (Ding et al., 2023) applied XGB-GP (XGBoost with Genetic Programming) to FDP for Chinese A-share listed companies on the Shanghai Stock Exchange. It examines the impact of pre- and post-COVID conditions on FDP, with a focus on feature importance and the role of genetic programming in enhancing result interpretation. The research highlights XGB-GP's effectiveness in adapting to pre- and post-COVID conditions, improving feature selection, and interpreting results. Alternatively, (Qian et al., 2022) in their study examine feature importance using a heuristic algorithm based on permutation importance (PIMP) with XGBoost for the same companies. By handling missing values with mean completer, PIMP refines feature selection and corrects biased importance scores, enhancing model accuracy. The research found that PIMP-XGBoost achieved the highest performance improvements, with ensemble models also benefiting from PIMP integration.

Shifting focus from XGBoost, (Kim and Upneja, 2021) in their study focus on predicting restaurant business failures by developing an ensemble model using Decision Trees (DT) based on voting mechanisms. By combining DT and logistic regression (logit) as base algorithms, and employing winsorizing for outlier handling, the study identifies key indicators such as low Operating Cash Flow and Accrued Income Debt (OCFAID), high KZ index, and low stock price as predictors of

failure. The ensemble model enhances prediction accuracy and interpretability, demonstrating that different financial and market-driven variables become significant depending on the economic context. The integration of multiple models further enhances predictive accuracy by leveraging the strengths of each individual algorithm. Considering this context, (Wang and Chi, 2024) in their research focus on Chinese listed companies from 2000 to 2020, utilizing a cost-sensitive stacking (CSS) method. The CSS method improves the identification of distressed firms up to five years ahead by combining cost-sensitive learning with stacking. The study highlights liquidity ratios as crucial predictors and shows that CSStacking outperforms methods like CSLR, CSDT, CSRF, and CSXGBoost in recall, AUC, G-mean, and Type II error.

2.4.3.1 Feature Selection and Preprocessing

Broadening the scope in areas of preprocessing, (Tsai et al., 2021) evaluated 10 datasets, comparing feature selection algorithms (Information Gain, Genetic Algorithm, Discriminant Analysis, Logistic Regression, and Particle Swarm Optimization) and ensemble methods (bagging and boosting). The study found that Genetic Algorithm (GA) was the most effective for feature selection. Ensemble methods generally outperformed individual classifiers, with the best performance achieved by combining affinity propagation with ensemble techniques. Enhancing the discussion on feature selection, (Ben Jabeur and Serret, 2023) introduces a novel forecasting approach for financial distress by combining Fuzzy Set Qualitative Comparative Analysis (FSQCA) with Convolutional Neural Networks (CNN), termed FCNN. Using data from French firms that filed with commercial courts between 2014 and 2016, the study evaluates feature selection methods such as t-test, stepwise discriminant analysis, stepwise logistic regression, and partial least squares discriminant analysis. The FCNN model significantly surpasses traditional methods like neural networks, logistic regression, and support vector machines in accuracy and error rates, especially for short-term predictions.

Shifting the focus to the other aspects of preprocessing and modeling, (Elhoseny et al., 2022) introduces the OD-PODNN model, using data from Australian, AnalCat, and Polish companies. This model combines outlier detection with Isolation Forest, classification via a Deep Neural

Network (DNN), and hyperparameter tuning through the Political Optimizer (PODNN). The preprocessing steps include imputing missing values with mean or mode and conducting feature selection to remove redundant features. Isolation Forest detects outliers, which are then classified using the DNN, with hyperparameters fine-tuned by the Political Optimizer. Focusing on optimization and fine-tuning aspects, in this research, (Xiao et al., 2024a) develops a framework for enterprise risk assessment and prediction using data from SMEs listed on the Shenzhen Stock Exchange between 2010 and 2020. This framework leverages OPTUNA optimization to enhance the LightGBM model's performance. The study also incorporates combined weighting and game theory to derive Financial Risk Value (FRV), which significantly improves prediction precision, indicator relevance, and overall model performance.

2.4.3.2 Usage of SVM

Support Vector Machines (SVM) are preferred for financial distress prediction due to their ability to handle complex, high-dimensional datasets effectively. SVM excels in identifying key features and managing non-linear relationships through kernel functions, making it well-suited for detecting subtle distress signals.

In that regard, A recent study by (Yang, 2023) introduced the SRA-SVM model, which combines Support Vector Machine (SVM) with Stepwise Regression Analysis (SRA) for feature selection, effectively addressing multicollinearity. Using data from Chinese companies from 2016 to 2021, the model showed superior performance, particularly when utilizing continuous financial ratios over individual ones. The SRA-SVM model excelled in accurately predicting financial distress within two years before the event, though its accuracy declined when predicting five years in advance, likely due to financial statement distortions under pressure. Developing the topic further, (Zeng and Yang, 2020) introduced a hybrid method that combines Sparse Principal Component Analysis (SPCA) with Support Vector Machines (SVM) to improve prediction for Chinese listed companies. This approach tackles the issue of high-dimensional financial data by employing SPCA to reduce dimensionality and eliminate redundant information, thus simplifying the dataset while

retaining key features. The SPCA-SVM method outperformed models using all original indicators, especially in decreasing the misclassification of distressed companies as financially stable.

Concentrating on the capabilities of ensemble methods and unlabelled data, (Chen et al., 2020) addresses prediction using Learning with Label Proportions (LLP), an approach ideal for situations with limited labeled data. The study utilizes proportion SVM methods, specifically bagged pSVM and boosted pSVM, to enhance model accuracy. Bagged pSVM improves robustness by combining multiple SVM models trained on different data subsets, while boosted pSVM focuses on refining predictions for harder-to-classify cases. Employing it to its best advantage, through this study, (Liang et al., 2020) uses a stacking ensemble method to integrate diverse data sources, with Support Vector Machines (SVM) acting as both base and meta learners. This approach is critical as governance factors often impact a firm's financial health and likelihood of distress, offering insights beyond those provided by financial ratios alone.

2.4.3.3 Other Advanced Ensembles

This study carried out by (du Jardin, 2021b) utilizes data from the Diane database, focusing on French firms' balance sheets and income statements, to improve prediction through a novel bi-clustering approach. By partitioning firms into homogeneous groups based on shared financial patterns and applying self-organizing maps to classify these groups, the study addresses the complexity and variability inherent in financial data. This bi-clustering method enhances the model's ability to capture nuanced variations within and between clusters, offering a significant improvement over traditional models that struggle with high-dimensional data. Later building on the framework, the authors (du Jardin, 2021a), explore clustering firms based on their structural patterns and growth trajectories over time. Their research employs ensemble methods like bagging and boosting, using Decision Trees (DT) as base classifiers, and integrating self-organizing neural networks to address clusters with similar financial patterns.

In another regard, (Rahayu and Suhartanto, 2020) uses a different approach by employing Case-Based Reasoning (CBR) ,by retrieving similar historical cases based on financial ratios and reusing class labels through a revise-retain process. While CBR may not match the accuracy of advanced machine learning models, it excels at identifying distressed firms, making it effective for intelligent financial distress prediction. Further delving into the topic of explainable AI, (Zhang et al., 2022) explores AI for prediction using data from Chinese listed companies (2007-2020). The research balances model accuracy with interpretability by using feature selection techniques and SHAP (SHapley Additive exPlanations) and counterfactual explanations for clarity. It evaluates ensemble models like LightGBM, XGBoost, and Random Forest, alongside base classifiers such as Logistic Regression, Support Vector Machine, and Decision Tree. LightGBM achieved the highest AUC of 0.92. SHAP and partial dependence plots (PDPs) are used for feature importance and interaction visualization, respectively, enhancing model interpretability.

2.4.4 Integrated Prediction Methods

2.4.4.1 Text Analysis

Integrated prediction methods combine diverse data sources, such as text and network data, to enhance predictive accuracy and depth. By merging unstructured text (like news, financial reports, social media sentiment) with network data (reflecting relationships and interactions), these methods offer a comprehensive view of influencing factors. This synergy uncovers complex patterns often missed in separate analyses, benefiting areas like financial distress prediction, fraud detection, and forecasting. Leveraging both data types, integrated methods improve prediction robustness, adaptability, and precision, leading to better decision-making.

Delving into additional data possibilities, the study undertaken by (Jiang et al., 2023), presents an innovative approach by integrating semantic features derived from patent texts with traditional accounting data. Their research highlights the significance of non-financial indicators, specifically technological attributes such as novelty and beneficial effects, which are often overlooked in conventional models. Widening the perspective, (Zhao et al., 2022) improves prediction by integrating sentiment tone features from online stock forums, management discussion and analysis

(MD&A), and financial statement notes (FSN) with traditional financial data. The study finds that recent comments on stock forums notably enhance FDP accuracy due to their relevance and timeliness. Using CatBoost, the research outperforms models like Logistic Regression (LR), Decision Trees (DT), Support Vector Machines (SVM), XGBoost, and Artificial Neural Networks (ANN). As an alternative, the textual data sourced from 10-K filings and management discussion & analysis reports is utilized by (Mai et al., 2019) in their research. The research reveals that deep learning models, especially those combining deep learning with average embedding techniques, significantly improve performance with textual data compared to numerical (accounting) data alone.

Expanding the scope with advanced models, this study leverages unstructured data from auditors' reports and management statements of Danish non-financial and non-holding private limited and stock-based firms. The authors (Matin et al., 2019), use Convolutional Neural Networks (CNNs) for initial pattern extraction and Recurrent Neural Networks (RNNs) with attention mechanisms for deeper understanding of textual data. CNNs identify complex patterns, while RNNs with attention focus on relevant information for better predictions. The study finds that including auditors' reports significantly improves model performance, increasing the Area Under the Curve (AUC) by 2% compared to models using only management statements.

On the other hand, (Huang et al., 2024) developed the Variational Deep Financial Distress Prediction (VDFDP) method, utilizing diverse data sources from the Chinese market, including financial indicators, annual reports, stock forums, legal judgments, and financial news. The VDFDP model incorporates encoder and view fusion techniques to address the challenges of incomplete multi-view data. The approach integrates SMOTE and focal loss to manage class imbalance. These advanced techniques together enhance the model's accuracy and reliability across various data sources. Integrating BERT, (Hajek and Munk, 2024) developed a hybrid model using BERT for sentiment analysis of annual reports and integrating XGBoost with unsupervised learning. This approach handles class imbalance and outlier detection robustly, combining BERT's nuanced sentiment analysis with financial indicators.

2.4.4.2 Network Analysis

Network data is increasingly important in finance, revealing insights that traditional metrics may miss by analyzing relationships between entities like businesses, investors, and markets. In financial distress prediction, network data highlights patterns of interconnectedness that indicate systemic risk and influence. Integrating this data into machine learning models enhances the understanding of these dynamics, leading to more detailed and comprehensive risk assessments.

Extending that discussion, (Kadkhoda and Amiri, 2024) proposes a hybrid model for predicting financial distress by integrating network analysis with machine learning. This approach uses company networks based on financial indicator similarity and correlation to capture inter-firm relationships. By combining network-centric features like Closeness Centrality with traditional financial ratios, the model enhances predictive accuracy, revealing complex interdependencies among financial entities. This study undertaken by (Wang et al., 2024b), employs graph-based learning techniques, specifically Graph Neural Networks (GNNs), to analyze the impact of various events that are both positive and negative on financial distress. It uses data from non-listing companies from the National Equities Exchange and Quotations (NEEQ) in China. The NetRisk framework, which incorporates the Adaptive Interpretable Graph Contrastive Learning (AIGOL) model, enhances the assessment of financial distress for small and medium-sized enterprises (SMEs) by integrating network and event data. AIGOL improves prediction accuracy by evaluating the influence of indirect events and generating network embeddings, providing a more nuanced analysis than traditional financial metrics and baseline methods.

Expanding the comprehension and investigating the potential of GNN, (Wang et al., 2024c) introduces a "represent-then-discover" framework that leverages network data to capture heterogeneous interactions among entities. Two novel models are proposed, namely the Entity-Importance Graph Contrastive model (EIGC) and the Subgraph-Distillation Frequent Interaction Mining model (SDIM). EIGC focuses on entity-importance aware sampling, heterogeneous information aggregation, and uses a canonical correlation analysis-based loss function to address

the unbalanced distribution of entities. SDIM identifies key interaction patterns such as multiple interactions, chain structures, and investment-management interactions.

2.4.5 Advanced Models for Financial Distress Prediction

Recent studies blend these methods and, at times, introduce novel advanced approaches for prediction.

In connection with that, (Che et al., 2024) investigates predictive capabilities for companies listed on China's National Equities Exchange and Quotations (NEEQ) from 2019 to 2021 using a multi-modal approach. The model integrates financial indicators, interim networks, and current reports, employing modality-specific attention mechanisms to emphasize relevant features from each data source. To address class imbalance, the study incorporates focal loss, adjusting sample weights during training to reduce the impact of imbalanced data. This methodology enhances the model's handling of diverse data types and improves overall predictive accuracy.

In a different vein, addressing the automation potential, this study carried out by (Papík and Papíková, 2024), delves into the application of automated machine learning (AutoML) for predicting bankruptcy in manufacturing companies, using data from 2020 to 2021 and focusing on Industry 4.0. It compares several AutoML frameworks and finds that H2O's algorithm outperforms others, including CatBoost and XGBoost, in terms of predictive accuracy. The research highlights AutoML's advantages in streamlining the modeling process, which is particularly beneficial for manufacturing companies. However, it also notes that performance can vary significantly between different AutoML algorithms. The study underscores the need for ongoing exploration of additional algorithms, such as TPOT and AutoWEKA, to further enhance predictive capabilities.

Building upon the research on automation potential, (Balachander et al., 2023) introduced the FCPFS-QDNN technique for automated financial crisis prediction. This innovative approach combines feature subset selection (FS) and quantum deep neural networks (QDNN) to enhance

predictive accuracy. The method normalizes financial data into a scalar format, employs the ISA-FS method for relevant feature subset selection, and applies the QDNN model for prediction. Experimental results reveal that integrating feature selection with machine learning markedly improves the predictive performance of the FCPFS-QDNN technique.

2.5 Data Quality and Challenges

Data quality is crucial for accurate financial distress prediction (FDP), as the reliability of predictive models hinges on the integrity of the data. Incomplete or inaccurate data can lead to flawed predictions, making it essential to implement robust data preprocessing techniques. Advanced methods are employed to address these challenges. However, due to regulatory constraints, the complexity of financial data, and its dynamic nature, maintaining data quality remains a persistent challenge in the finance sector.

2.5.1 Class Imbalance

Since the proportion of financially healthy companies is consistently higher compared to distressed ones, class imbalance becomes a significant issue, especially when working with real data.

Exploring the context, (Cheng et al., 2021) addresses the challenges of imbalanced data and missing values in financial fraud prediction by leveraging data from SFIPC and TEZ, specifically focusing on identifying financial statement fraud. The dataset includes financial statements from companies involved in financial fraud. To manage data imbalance, they employ random oversampling and SMOTE, comparing these techniques to understand their efficacy. The results indicate that oversampling methods outperform undersampling approaches in dealing with class imbalance. For missing values, they utilize listwise and pairwise deletion methods, opting to remove instances with missing data to maintain dataset integrity.

Further broadening the application scope of SMOTE, (Garcia, 2022) examines bankruptcy prediction for U.S. firms, focusing on class imbalance and the efficacy of SMOTE (Synthetic Minority Over-sampling Technique). The study shows that combining SMOTE with cluster-based undersampling achieves the best classification performance, improving recall, precision, specificity, G-mean, F-measure, and AUC. It highlights that machine learning techniques, particularly with SMOTE, outperform traditional methods. The research emphasizes the importance of recall, noting that SMOTE and its variants, such as ADASYN, DB SMOTE, and SMOTE-CBU, enhance recall accuracy, with SMOTE-CBU providing a good balance of efficiency and performance. Random forest emerges as the top-performing classifier when paired with SMOTE or its variants.

(Veganzones and Séverin, 2018) in their study, by Investigating through a comparative analysis of various methods, assesses the performance of bankruptcy prediction models on imbalanced datasets. They compare oversampling, undersampling, SMOTE, and EasyEnsemble and tested it with models including Logistic Regression, Support Vector Machines, Neural Networks, and Linear Discriminant Analysis, using financial ratios and other variables. Key findings indicate that imbalances, especially when bankrupt firms make up 20% or less of the sample, significantly degrade model performance. SVM is the least affected by moderate imbalances but shows substantial performance loss in extreme cases (90/10 and 95/5 class proportions). Among the techniques, SMOTE outperforms others in improving prediction accuracy across various imbalances and dataset sizes. Oversampling methods generally restore 43.9% of performance on average, but their effectiveness diminishes as dataset size grows, eventually plateauing.

Similarly, this study by (Alam et al., 2021) explores corporate bankruptcy prediction using machine learning, addressing the challenges of imbalanced datasets with the Polish bankruptcy dataset. Missing data is imputed with median values. Various data balancing techniques, including random undersampling and SMOTE, are applied, with SMOTE showing better performance by generating synthetic samples and preserving information. Among the classifiers tested such as SVM, LMT, J48, Random Forest, and Decision Forest. The Decision Forest (with 128 trees and

depth of 32) achieves superior accuracy. Z score normalization is used to standardize data, with Decision Forest consistently outperforming other models.

(Du et al., 2020) leverages data from the China Stock Market and Accounting Research database, spanning January to December 2018, to address FDP for 3,670 publicly listed companies on the Shanghai and Shenzhen stock markets. It proposes an ensemble approach combining cluster-based undersampling with Gradient Boosting Decision Trees (GBDT) and XGBoost to overcome class imbalance issues. The research evaluates five feature selection methods—LASSO FS, tree-based FS, L1-based FS, f_classif FS, and XGBoost-based FS—to refine the dataset, selecting the most relevant features for the models. The CUS-GBDT model, integrated with XGBoost, significantly outperforms traditional models, demonstrating superior predictive accuracy and efficiency.

Following a similar approach on comparing ensembles and imbalance techniques, (Antulov-Fantulin et al., 2021) aims to predict the bankruptcy of Italian municipalities from 2009 to 2016 using financial, socio-demographic, and economic data. The study employs Gradient Boosted Machines, Random Forest , Lasso regression, and Neural Networks, addressing class imbalance with class weights, undersampling, and random oversampling. Undersampling with GBM yields the highest accuracy, ROC, and PRC scores. Analyzing 7,795 municipalities, the study underscores the significance of non-financial features—such as regional indicators and socio-demographic characteristics—as well as financial metrics like deficits and debt ratios in predicting municipal defaults. Analyzing the integration opportunities of SMOTE, (Shen et al., 2020) presents the Adaptive Neighbor SMOTE-Recursive Ensemble Approach (ANS REA), a novel model that improves class imbalance handling through adaptive sampling and ensemble techniques. The Dynamic Financial Distress Prediction (DFDP) method, introduced in their study, dynamically updates financial distress forecasts to address concept drift in continuous data streams. DFDP outperforms other methods, including SMOTE, ANS, RWO, Racog, SMOTEBoost, SMOTEBagging, and Majority Weighted Minority Over-sampling (MWMOTE).

Instead of traditional methods like SMOTE or oversampling, (Zoričák et al., 2020) tackles imbalance through alternative approaches. Using financial ratios as input features, the study applies one-class classification techniques, which utilize only samples from the majority class to train the model. The techniques explored include One-Class SVM, Least-Squares Anomaly Detection (LSAD), and Isolation Forests. Among these, LSAD is identified as the most effective model for predicting fiscal strain. The research demonstrates that one-class classification models are particularly useful for distinguishing financially distressed companies. (Zhou et al., 2021) integrates four datasets, including Chinese data, to evaluate advanced feature selection, class imbalance handling, and classifier combinations. SMOTE is used to address class imbalance by generating synthetic samples, enhancing model performance. LASSO (Least Absolute Shrinkage and Selection Operator) addresses multicollinearity in financial distress datasets by selecting relevant features and shrinking less important coefficients. This method improves model robustness and accuracy, particularly for imbalanced datasets. The study finds that SVM (Support Vector Machine), combined with LASSO, performs exceptionally well in financial distress prediction, effectively handling both balanced and imbalanced datasets and defining clear decision boundaries.

Concentrating on other challenges, (Nyitrai and Virág, 2019) investigates methods for handling outliers in a Polish bankruptcy dataset, assessing their impact on model performance. Four strategies are compared such as raw data without outlier treatment, winsorization (which caps extreme values), outlier deletion, and CHAID decision tree categorization. CHAID decision trees, which segment data into categories based on the most significant splits, are found to be the most effective in managing outliers, offering robust performance despite the presence of outliers. In contrast, decision trees themselves are noted for their resilience to outliers, making them a reliable choice for robust predictions. On the other hand, artificial neural networks (ANN) and linear models show sensitivity to outliers, leading to potential distortions in prediction accuracy.

Expanding on the issue of imbalance , (Xiao et al., 2024b) employs a combination of Variational Autoencoder (VAE) and Deep Forest Ensemble (DF) to tackle class imbalance and improve model interpretability. The VAE-DF model demonstrates effectiveness in handling highly imbalanced and

non-linear data. VAE-based oversampling outperforms traditional methods such as Random Oversampling, SMOTE, ADASYN, and other SMOTE-related models. This approach enhances performance by generating synthetic samples that better capture the underlying data distribution, leading to more accurate and interpretable credit scoring predictions. The results highlight the VAE-DF model's superiority in managing class imbalance and its robustness in predicting credit risk. Delving deeper, (Al Ali et al., 2024) presents an advanced method for managing class imbalance and feature selection. The approach uses the Hamann Similarity Indexed Chinese Whispers clustering process to group data samples based on similarity indices, which helps mitigate class imbalance by balancing data across clusters and reducing bias towards the majority class. The study introduces the Chinese Whisper Clustered Stochastic Gradient Descent Federated Learning (CWCSGDFL) method, which clusters similar samples to manage imbalance effectively and streamline training, thus reducing time complexity. For feature selection, the Kaiser-Meyer-Olkin (KMO) correlative targeted projection model is applied to select the most relevant features, enhancing predictive accuracy.

2.5.2 Feature Importance and Imputation

Financial datasets often contain high-dimensional data from sources like transactions, financial statements, and market data, which can lead to missing or incomplete information. This poses challenges for machine learning models and may reduce prediction accuracy. Imputation is essential for filling these gaps, minimizing bias, and enhancing model performance. Feature selection is equally important in managing high-dimensional data, as it identifies and retains the most relevant variables, simplifying models, reducing overfitting, and improving computational efficiency. Combining effective imputation with careful feature selection allows models to handle missing data and focus on critical information, resulting in more accurate and reliable financial predictions.

Considering this problem, (Ben Jabeur et al., 2023) introduces FS-XGBoost, a novel approach for bankruptcy prediction that combines feature selection with XGBoost. The method employs stepwise discriminant analysis, stepwise logistic regression, and partial least squares discriminant

analysis (PLS-DA) for feature selection. FS-XGBoost outperforms traditional techniques, showing improved discrimination power and accuracy, particularly in AUC. By reducing the number of prediction variables, FS-XGBoost decreases processing time and model complexity. The study also addresses class imbalance through oversampling, enhancing the model's overall effectiveness. FS-XGBoost's integration of feature selection with XGBoost makes it a highly efficient tool.

(Yu and Li, 2023) examines methods for handling missing data in financial datasets of China's A-share listed enterprises, focusing on multiple imputation by chained equations (MICE) using random forest (RF) for imputation. The study finds that MICE with RF effectively addresses the challenges posed by substantial amounts of missing data, although it is noted for its computational intensity and time consumption. MICE with RF performs well when data are missing at random, proving to be a robust method for imputation in such scenarios. Additionally, the study introduces a case-based reasoning (CBR) driven imputation method, which also demonstrates high performance. Despite the benefits of CBR, MICE with RF remains a capable and reliable model, particularly in handling large datasets with missing values. Further delving deep in exploring investigating MICE, the study by (Samad et al., 2022) explores enhancing missing data imputation and classification accuracy by integrating Deep Neural Networks (DNN) with Multiple Imputation by Chained Equations (MICE). Evaluating various datasets, including credit card default, breast cancer, and dermatology, the study focuses on MICE's imputation performance with the credit card default dataset. By replacing MICE's traditional linear regressor with a DNN, the study aims to improve imputation accuracy and classification performance. The DNN's ability to capture complex data patterns leads to more precise imputation, significantly improving data quality and predictive model accuracy across different domains.

(Hassan and Yousaf, 2022) examines bankruptcy prediction using a Polish company dataset, focusing on preprocessing and handling data imbalance. The study addresses missing data and class imbalance through various imputation techniques and oversampling methods, finding that simple mean imputation outperforms more complex methods. Nine classification algorithms, including Gaussian Naïve Bayes, logistic regression, and random forests, are modeled. The balanced bagging classifier with mean imputation achieves the highest accuracy of 98.23%,

surpassing other models. SMOTE is used to balance the training dataset, and 36 analyses with k-fold cross-validation evaluate model effectiveness.

As observed earlier with CBR, and further examining its capabilities on imputation, (Yu et al., 2024) investigates imputation methods for missing data in the context of Chinese A-share listed enterprises, focusing on both single and multiple imputation techniques. It introduces a novel hybrid approach combining CBR with weighted imputation, aimed at improving the accuracy and reliability of missing data handling. Additionally, the study proposes the Learning Vector Quantization-Case-Based Reasoning (LVQ-CBR) model to address class imbalance issues

Recent research has made significant strides in both imputation and class imbalance techniques, but the exploration of their combined effects remains limited. While each method has advanced individually, the impact of these techniques on preserving the originality of the data when used together is still underexplored.

This gap highlights the need for more studies focusing on how these methods interact and how they can be refined to maintain the integrity of the original data while improving model performance.

2.6 Discussion

Extensive research in financial distress prediction has focused on data preprocessing, model development, and optimization, evolving from basic to advanced techniques with the advent of machine learning. Researchers have explored various data sources to improve predictive accuracy and reliability. Despite these advances, significant challenges remain, particularly due to the nature of financial data. Real-world datasets often suffer from class imbalance, with far fewer companies in distress than those stable, skewing predictive models toward the majority class. Additionally,

issues like missing data, outliers, and inconsistencies complicate the prediction process without distorting genuine data patterns.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This study concentrates on the implementation and evaluation of various imputation and imbalance handling techniques, assessing their impact on financial distress prediction. This section provides an overview of the methods and metrics employed in the research, offering a detailed exploration of the rationale behind these choices and their significance in the overall analysis.

3.2 Algorithms and Techniques

3.2.1 Exploratory analysis

This section discusses the methodologies used to explore and analyze the dataset for predicting financial distress. The methods employed primarily focus on understanding the data's nature, relationships, and patterns. These discussed techniques are employed for EDA and explained in Section 4.3.

3.2.2 Isolation Forest

The authors, (Zoričák et al., 2020; Elhoseny et al., 2022; Hajek and Munk, 2024)have Implemented isolation forests for outlier detection and management. Outliers can significantly affect the performance of predictive models, making their identification crucial. Isolation Forest works by isolating observations in a dataset and is particularly effective because it focuses on isolating anomalies rather than profiling normal data points. By identifying and handling these outliers, the model’s accuracy and reliability are improved. This method was selected as it effectively detects outliers that are difficult to identify in high-dimensional data, particularly those with complex patterns in this study.

3.2.3 Z-score

The Z-score method is also utilized for outlier detection. The Z-score measures the number of standard deviations a data point is from the mean of the dataset. Data points with a Z-score beyond a certain threshold (commonly ±3) are considered outliers. This method is effective in identifying extreme values that could distort analysis or model performance, particularly by detecting points that deviate significantly from average behavior. It complements the Isolation Forest approach, making it a valuable addition to this study.

3.2.4 Pearsons

Pearson's correlation coefficient is used to measure the linear relationship between pairs of variables in the dataset. This coefficient ranges from -1 to 1, where values close to 1 indicate a strong positive linear relationship, values close to -1 indicate a strong negative linear relationship, and values around 0 suggest no linear relationship. Understanding these relationships is crucial for identifying patterns and choosing the most suitable methods and models for prediction, making this approach a key choice in this study.

3.2.5 Spearman

Spearman's rank correlation is employed to assess the monotonic relationship between variables, which may not necessarily be linear. Unlike Pearson's correlation, Spearman's does not assume a normal distribution of the data and is better suited for capturing non-linear relationships. By using both Pearson's and Spearman's correlations and assessing their difference, the study can distinguish between linear and non-linear patterns in the data, providing a more comprehensive understanding of the relationships between variables.

3.2.6 Kurtosis

Kurtosis is used to understand the data distribution. It measures the extremity of data points in the tails relative to a normal distribution. High kurtosis indicates heavy tails, meaning more data points

are found in the tails, which could suggest the presence of outliers. Conversely, low kurtosis suggests lighter tails.

3.2.7 Skewness

Skewness is a measure of the asymmetry of the data distribution. A skewness value of zero indicates a perfectly symmetrical distribution, while positive or negative values indicate right or left-skewed distributions, respectively.

3.2.8 Correlation matrix – heatmap

To visualize the relationships between all variables simultaneously, a correlation matrix is constructed and represented as a heatmap. A correlation matrix contains the Pearson correlation coefficients for all pairs of variables in the dataset. This visualization aids in understanding the overall structure of relationships within variables in the data, highlighting potential multicollinearity issues.

3.2.2 Data Preprocessing

This section details the methods implemented in this study for imputing null values and addressing class imbalance.

3.2.2.1 Imputation

The dataset contains interconnected variables that reflect the financial outlook of the companies, with ratios derived from balance sheets and income statements. In such cases, where columns are closely related, missing values in one column cannot be imputed based on the column alone, instead, the context of the entire row must be considered. To address this complexity, two advanced imputation methods are employed such as Multiple Imputation by Chained Equations (MICE) and Autoencoders. These techniques are particularly effective in maintaining the integrity of the data by considering the interdependencies between variables during the imputation process.

3.2.2.1.1 Multiple Imputation by Chained Equations (MICE)

MICE is a statistical technique that handles missing data by creating multiple complete datasets through an iterative process. In this study, MICE is used by incorporating Random Forest Regressor as the predictive model, given the non-linear nature of the dataset. MICE is selected due to its robustness and flexibility in handling missing data by generating multiple imputed datasets. This method is particularly useful in capturing the uncertainty associated with missing values. (Yu and Li, 2023)suggests that incorporating Random Forest Regressor enhances MICE's effectiveness by leveraging its ability to model complex, non-linear relationships between variables.

$$X_j^{t+1} = f_{j\left(X_{-j}^{t}, \theta_j^{t}\right)}$$

This formula represents the iterative process of MICE. For each feature j with missing values, it estimates those values based on all other features (X_j) using a Random Forest model (f_j). This process repeats for multiple iterations (t) until convergence.

$$y_{pred} = \frac{1}{n_{trees}} * \Sigma\left(tree_{prediction}\right)$$

The random forest used in each step of imputation is an ensemble of decision trees, denoted by the formula given above.

By generating multiple imputed datasets, MICE facilitates a more accurate estimation of missing values and their effects on the final analysis. This technique has proven effective in addressing robust missing data challenges, as demonstrated in studies by (Samad et al., 2022; Yu et al., 2024). However, it is important to note that studies indicate MICE can be time-consuming.

3.2.2.1.2 Autoencoders (AE)

Autoencoders are employed as a sophisticated technique for imputing missing values as it is generally suitable for adapting and handling high-dimensional datasets. This type of artificial neural network is designed to learn efficient representations of the data. They consist of an encoder, which compresses data into a lower-dimensional latent space, and a decoder, which reconstructs the data from this representation. During training, the autoencoder minimizes reconstruction error,

refining the imputation of missing values. This method excels in capturing non-linear relationships in high-dimensional data that traditional methods might miss.

Formula explanation:

Encoder: $\sigma(W_e * x + b_e)$

Decoder: $x' = W_d * h + b_d$

Loss function: $MSE = \frac{1}{n} * \Sigma(x_i - x_i')^2$

This Encoder compresses the input data (denoted as x) into a lower-dimensional representation (denoted as h). It does this using a weight matrix (W_e) and a bias vector (b_e), and applies a non-linear activation function (σ), ReLU, to capture complex patterns in the data. Essentially, the encoder transforms the input into a more compact form that retains its essential features. The decoder then takes this compressed representation (h) and attempts to reconstruct the original input (x'). It uses another weight matrix (W_d) and a bias vector (b_d), but without a non-linear activation function, resulting in a linear transformation.

The effectiveness of the autoencoder is evaluated by comparing the reconstructed data (x') with the original data (x). This is measured using the Mean Squared Error (MSE), which calculates the average squared difference between each element of the original and reconstructed data. The goal during training is to minimize this MSE, ensuring that the autoencoder learns to produce reconstructions that are as close as possible to the original inputs.

Autoencoders effectively reconstruct the input data, filling in missing values while maintaining the essential patterns and relationships between interconnected variables, which is vital for this dataset. While studies (Lall and Robinson, 2022; Xiao et al., 2024b) use Autoencoders for class imbalance handling and dimensionality reduction by integrating them with other techniques, they demonstrate the capability of Autoencoders in capturing non-linear patterns in the data.

3.2.2.2 Class Imbalance

To address the class imbalance in the data, this study explores three methods such as ADASYN, KMeans SMOTE, and SVMSMOTE. The selection of these methods was motivated by the necessity to tackle the distinctive challenges inherent in financial datasets, including the management of noisy, complex, and highly imbalanced data. These models were chosen to provide a thorough approach and understanding to addressing imbalance.

3.2.2.2.1 ADASYN

ADASYN is an extension of SMOTE (Synthetic Minority Over-sampling Technique) that focuses on generating synthetic samples for minority class instances that are more challenging to classify. It adaptively selects the samples that are difficult to learn and generates more synthetic examples around these hard-to-classify instances. The dataset as observed to be non-linear, makes ADASYN suitable to generate meaningful synthetic samples, which may be difficult for traditional oversampling methods.

As observed in studies(Xiao et al., 2024b), ADASYN is robust in handling class imbalance on non-linear datasets. Utilizing a weighted distribution for various minority class instances based on their degree of learning difficulty, (Shen et al., 2020) show in their research the advantages of ADASYN.

3.2.2.2.2 KMeansSMOTE (KMSMOTE)

KMeansSMOTE enhances the original SMOTE algorithm by incorporating K-Means clustering, to address the limitations of uniform sample generation in the original SMOTE algorithm. It first clusters the minority class samples into several clusters using K-Means and then generates synthetic samples within each cluster. This method ensures that the synthetic samples are distributed more evenly across different regions of the feature space. While studies have not explored the use of KMeansSMOTE, (Veganzones and Séverin, 2018; Garcia, 2022) highlight the effectiveness of SMOTE and its variants in managing class imbalance.

KMeansSMOTE's capability to generate balanced and representative synthetic samples proves especially valuable in predicting financial distress, as it can identify and leverage clusters of companies exhibiting similar distress patterns. The targeted oversampling enhances the model's ability to learn the decision boundary more effectively by concentrating on critical cases that distinguish distressed companies from healthy ones.

3.2.2.2.3 SVMSMOTE

SVMSMOTE extends SMOTE by using Support Vector Machines (SVMs) to guide the synthetic sample generation process. It creates synthetic samples based on the decision boundary identified by an SVM classifier. This technique focuses on generating samples near the decision boundary, where class overlap is high, and helps to refine the separation between classes.

This method enhances model performance by generating more informative samples that improve class separability, particularly in financial distress prediction, where overlapping data from companies can be challenging to distinguish. (Garcia, 2022) in their study, examine the effectiveness of cluster-based SMOTE, particularly DB SMOTE, which generates synthetic samples along the shortest path from minority instances to cluster centroids. Building on a similar trajectory, this study explores the use of SVMSMOTE, using the SVM decision boundary to guide sample generation. This approach is particularly useful in financial distress prediction, where the boundary between distressed and non-distressed firms may be complex and non-linear.

3.2.2.3 Dimensionality reduction

3.2.2.3.1 Principal Component Analysis (PCA)

Given the high dimensional structure of this dataset, Principal Component Analysis (PCA) is employed as a dimensionality reduction technique. The high dimensionality of this data can pose significant challenges, including increased computational complexity, potential overfitting, and difficulties in model interpretability. PCA is a statistical method that transforms the original features into a smaller set of variables known as principal components. These components capture the maximum variance in the data, allowing for the reduction of dimensionality while preserving the essential information needed for predictive modeling.

Although numerous studies (Aydin et al., 2022; Qian et al., 2022) classify ratios under various headers, we employ Principal Component Analysis (PCA) for these groups to reduce dimensionality and preserve key data components. Prior research (Syed et al., 2023), has also highlighted specific ratios that significantly influence predictive performance. In this study, financial ratios are initially organized under respective headers based on their nature and relevance. PCA is then applied to each group to decrease dimensionality. Each group is specifically reduced to two principal components, ensuring that the most critical information is preserved while minimizing redundancy and noise.

By reducing the dimensionality of each group to two components, the complexity of the dataset is significantly decreased, facilitating more efficient model training and evaluation. These two components from each group are then combined, along with their respective headers, to create a consolidated feature set. This consolidated set is subsequently used to train and evaluate predictive models.

3.2.3 Model training

3.2.3.1 Artificial Neural Networks (ANN)

ANNs are inspired by the structure of the human brain and consist of layers of interconnected neurons (Brenes et al., 2022). Data is passed through these layers, with each neuron applying an activation function to transform the input. The network is trained using backpropagation, which adjusts the weights based on the error between the predicted and actual outcomes. ANNs are capable of modeling complex, non-linear relationships due to their multiple hidden layers and non-linear activation functions. Studies (Wu et al., 2022; Titikkristanti and Mahardika, 2023) have demonstrated that ANN has achieved higher accuracy than other non-linear techniques.

Their ability to model intricate patterns and interactions among variables makes them a powerful tool for capturing the underlying dynamics of financial distress in this research.

3.2.3.2 Support Vector Machine (SVM)

SVM is a supervised learning algorithm that constructs hyperplanes in a high-dimensional space to separate different classes. It is particularly effective in scenarios where the margin of separation between classes is small or non-linear. The use of kernel functions allows SVM to handle non-linear relationships by transforming the input space into a higher-dimensional space where a linear separation is possible. We use Linear and RBF kernel functions in this study. SVM's ability to find the optimal boundary between distressed and non-distressed firms, even in cases of class overlap, makes it suitable for this study.

Particularly, SVM variants and integrated SVM methods have demonstrated notable effectiveness in financial distress prediction due to their robustness as seen with studies (Huang and Yen, 2019; Zeng and Yang, 2020; Yang, 2023).

3.2.3.3 Logistic Regression (LR)

Logistic Regression (LR) is a widely used statistical method for binary classification problems. It models the probability that a given input belongs to a particular class by fitting a logistic function to the data. LR's interpretability is advantage, where understanding the influence of individual financial ratios on the probability of distress is crucial. It is particularly useful when the relationship between the predictors (financial ratios) and the outcome (financial distress) is approximately linear. LR has been a foundational tool in financial distress prediction, for its efficiency in binary classification, where its effectiveness has been amplified when integrated with other methods (Serrano-Cinca et al., 2019; Son et al., 2019).

Incorporating Logistic Regression (LR) alongside other non-linear techniques enriches the study, offering a more complete understanding of the predictive capabilities of various models.

3.2.3.4 Random Forest (RF)

Random Forest is an ensemble method that constructs multiple decision trees during training. Each tree is built using a random subset of features and data samples. It reduces the risk of overfitting by averaging multiple trees and considering only a subset of features at each split (Cheng et al.,

2021). Random Forest is employed in this study because it handles high-dimensional data and complex interactions well.

3.2.3.5 Decision Trees (DT)

Decision Trees are simple yet powerful models that split the data into subsets based on the value of input features, forming a tree-like structure. They are easy to interpret. However, they are prone to overfitting. Decision Trees (DT) are widely utilized classifiers in financial distress prediction, particularly in ensemble models, where they are often combined with other traditional metrics for enhanced predictive accuracy (Putri and Dhini, 2019; Kim and Upneja, 2021; Wang et al., 2022a).

3.2.3.6 Extreme Gradient Boosting (XGBoost/XGB)

XGBoost is another widely used model that is frequently combined with other advanced techniques due to its flexibility and powerful performance (Liu et al., 2022; Ding et al., 2023; Hajek and Munk, 2024). It is an advanced implementation of gradient boosting that is designed for speed and performance (Carmona et al., 2019). It builds decision trees sequentially, where each tree attempts to correct the errors of the previous one. XGBoost includes several regularization techniques to prevent overfitting and is capable of handling missing data and outliers effectively.

Its ability to model non-linear relationships and handle imbalanced data is critical in this study, ensuring high predictive performance across various scenarios.

3.2.3.7 Gradient Boosting Machine (GBM)

Gradient boosting machine (GBM) is an ensemble technique that builds models sequentially by fitting new models to the residuals of previous models. It focuses on reducing the prediction errors iteratively, which leads to improved accuracy.

GBM's flexibility and ability to improve performance through iterative learning are valuable in this study. It excels in situations where capturing subtle patterns in the data is crucial for making accurate predictions, particularly when the data contains complex interactions among financial ratios.

Its application alongside multiple sampling techniques and high dimensional data has been pivotal, highlighting its robustness in handling complex datasets (Antulov-Fantulin et al., 2021; Bitetto et al., 2023).

3.2.3.8 Naive Bayes (NB)

Naïve Bayes calculates the probability of each class given the features and selects the class with the highest probability. Despite the independence assumption, Naive Bayes often performs well in practice. NB offers a fast and simple model that can serve as a benchmark. As observed in studies (Hassanniakalager et al., 2021; Tsai et al., 2021), NB is recognized for its effectiveness when combined with other methods, its inclusion in this study enhances comprehensiveness by testing the capabilities of various models and algorithms.

3.2.4 Visualization

In this thesis, t-Distributed Stochastic Neighbor Embedding (t-SNE) and Pairwise Controlled Manifold Approximation Projection (PaCMAP) are employed as dimensionality reduction techniques for data visualization. These methods are particularly useful for understanding the structure of high-dimensional financial data.

3.2.4.1 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a non-linear dimensionality reduction technique that is particularly effective for visualizing high-dimensional data in 2D or 3D spaces. t-SNE works by minimizing the divergence between two distributions, one that measures pairwise similarities of the data points in the high-

dimensional space and another that measures these similarities in the low-dimensional space. It focuses on preserving the local structure of the data.

By visualizing the data, t-SNE helps in identifying the differences between each dataset and impact of each preprocessing technique.

3.2.4.1 Pairwise Controlled Manifold Approximation Projection (PaCMAP)

PaCMAP is another non-linear dimensionality reduction technique designed to preserve both local and global data structures. It is an improvement over t-SNE in terms of preserving the overall data distribution while still highlighting local clusters. By visualizing the data in three dimensions, PaCMAP helps in this study by identifying the differences in structure and pattern between each dataset and its preprocessing technique.

3.2.5 Evaluation

A range of metrics are employed to evaluate the datasets and the performance of the models. These methods also focus on understanding the impact of imputation and imbalance handling techniques. Each metric provides a unique perspective on model performance, offering a comprehensive understanding of the effects of these techniques.

3.2.5.1 Accuracy

Accuracy is a fundamental metric that measures the proportion of correct predictions made by the model out of all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

True Positives (TP) and True Negatives (TN) represent correct predictions. False Positives (FP) and False Negatives (FN) represent incorrect predictions. It provides a high-level, easily interpretable measure of the model's overall performance. However, in imbalanced datasets,

accuracy can be misleading, as a model may achieve high accuracy simply by predicting the majority class. Despite this limitation, accuracy remains an important baseline metric to track the general impact of imputation and class imbalance techniques on the model's predictive power.

3.2.5.2 Area under curve (AUC/ROC)

The AUC-ROC is derived from the ROC curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR).

$$\text{AUC} = \int_0^1 TPR(t)dFPR(t)$$

True Positive Rate (TPR) shows the proportion of actual positives correctly identified. False Positive Rate (FPR) shows the proportion of negatives incorrectly identified as positives. It evaluates the trade-off between true positive and false positive rates across all classification thresholds, making it particularly valuable when assessing models in scenarios with class imbalance. In this research, AUC-ROC is used to evaluate the overall discriminative capability of the models before and after applying imputation and class techniques.

3.2.5.3 Precision and Recall

Precision measures the proportion of true positive predictions among all positive predictions, reflecting the model's ability to avoid false positives. Conversely, Recall measures the proportion of actual positives that were correctly identified, indicating the model's ability to detect distressed firms.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

In this study, precision and recall are critical for balancing the trade-off between Type I errors (false alarms) and Type II errors (missed detections).

3.2.5.4 F1 Score

The F1 Score is the harmonic mean of precision and recall, providing a single, balanced measure that accounts for both false positives and false negatives.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

This metric is particularly relevant in financial distress prediction, where the goal is to minimize both types of errors.

3.2.5.5 Kappa

The Kappa statistic measures the agreement between the predicted and actual classifications, adjusting for the agreement expected by chance. This metric is crucial for evaluating models on imbalanced datasets, where high accuracy might be achieved by simply predicting the majority class.

$$\kappa = \frac{Po - Pe}{1 - Pe}$$

Observed Agreement (Po) is the proportion of times the model and actual classifications agree. Expected Agreement (Pe) accounts for agreement by chance. Kappa offers a more nuanced understanding of the model's true discriminative power and helps to reveal the effectiveness of the applied imbalance techniques.

3.2.5.6 G-mean

The G-mean metric evaluates the balance between sensitivity (true positive rate) and specificity (true negative rate). It is useful in scenarios with class imbalance, as it ensures that the model performs well across both the minority and majority classes.

$$G\ mean = \sqrt{Sensitivity \times Specificity}$$

In financial distress prediction, where sensitivity and specificity are crucial, the G-mean helps assess the overall balanced performance of the models, especially after applying imputation and class imbalance techniques.

3.2.5.7 Confusion Matrix

The confusion matrix provides a detailed breakdown of the model's performance by showing the number of true positives, false positives, true negatives, and false negatives. This granular information complements the other metrics by offering insights into specific areas where the model might be overfitting or underperforming, particularly in the context of class imbalance.

3.3 Methodology

3.3.1 End-to-End Pipeline

This segment encapsulates the pipeline of the data flow and the various processes involved, from data acquisition to the evaluation of predictive models. The Fig. 3.1 shows an overview of the flow of processes.

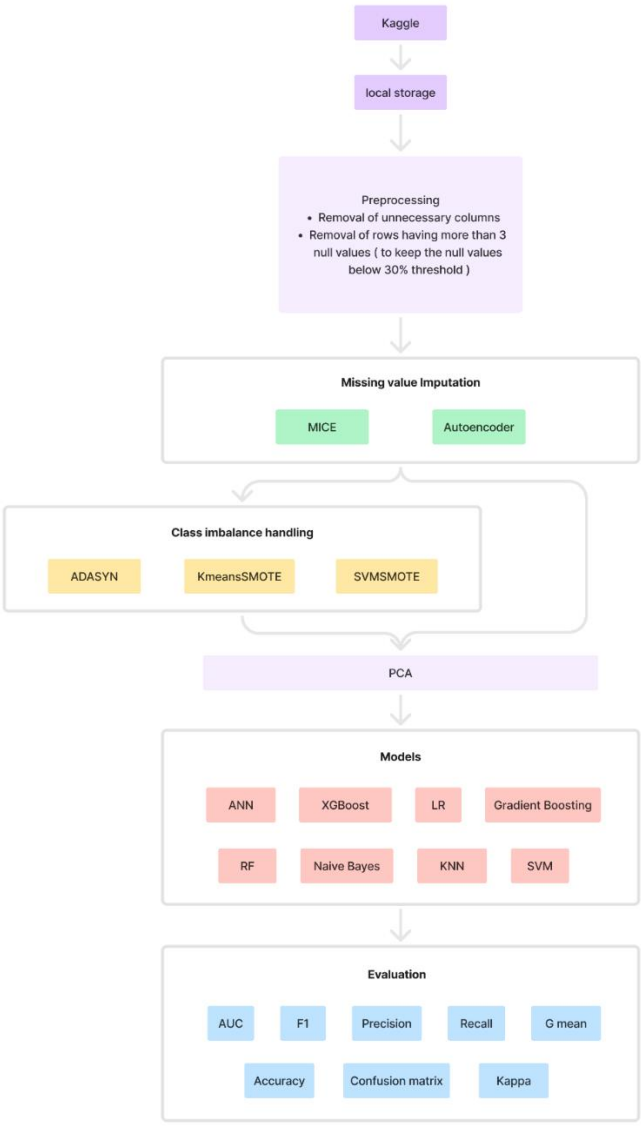

Figure 3.1 – End to End Pipeline

Each of the methods and techniques involved in various steps throughout the pipeline are comprehensively discussed in section 3.2.

3.3.1.1 Data acquisition

The data utilized in this study is sourced from Kaggle. The data is downloaded and stored locally for subsequent processing and analysis.

3.3.1.2 Preprocessing

Preprocessing is a critical step in the pipeline, aimed at refining the raw data to ensure its suitability for further analysis. The data sourced is then preprocessed. The preprocessing phase in this study involves:

- Removal of Unnecessary Columns: Since this study focuses on using ratios, other columns that do not contribute to the prediction task and are redundant are removed.
- Handling Missing Values: Rows with more than three null values are removed to maintain data quality. This criterion ensures that the dataset's integrity is preserved, keeping the proportion of null values below a 30% threshold.

3.3.1.3 Missing value imputation

Further the preprocessed dataset is then used for addressing the issue of missing values. The imputation techniques employed are:

- Multiple Imputation by Chained Equations (MICE)
- Autoencoder

These techniques result in the creation of two distinct datasets, one imputed using MICE and the other using Autoencoder.

3.3.1.4 Class imbalance handling

The derived two datasets are then used for addressing class imbalance. Each of the two datasets obtained from the imputation step undergoes the following class imbalance handling techniques:

- ADASYN
- KMSMOTE (KMeansSMOTE)
- SVMSMOTE

Applying these techniques to the two initial datasets results in the creation of six additional datasets, each balanced differently. This process ultimately produces a total of eight datasets, two with only imputation and six that are both balanced and imputed.

3.3.1.5 Principal component analysis (PCA)

Principal Component Analysis (PCA) is applied across all eight datasets to manage the high dimensionality. Ratios are grouped under their respective headers, and PCA reduces each group to two dimensions. These components are then combined, along with their headers, to form a consolidated data set.

3.3.1.6 Modeling

The reduced datasets are then utilized to train multiple machine learning models. The models considered include:

- Artificial neural networks (ANN)
- Support vector machine (SVM)
- Logistic regression (LR)
- Random forest (RF)
- Decision trees (DT)
- Extreme gradient boosting (XGBoost/XGB)

- Gradient boosting machine (GBM)
- Naive Bayes (NB)

These models are trained and evaluated using each of the eight datasets, ensuring a thorough exploration of the prediction potential across various approaches.

3.3.1.7 Evaluation

The performance of the models across various datasets is evaluated using a range of metrics, ensuring a comprehensive assessment:

- Accuracy
- Area under curve (AUC / ROC)
- F1 score
- Precision and recall
- Confusion matrix
- Kappa
- G-mean

These metrics are discussed in depth in section 3.2.6.

3.3.1.8 Summary

By systematically addressing the challenges inherent in financial data, such as missing values and class imbalance, and by employing a variety of models and evaluation metrics across multiple datasets, this methodology ensures robust and reliable predictive outcomes.

3.3.2 Data Selection and Description

This dataset contains real-world financial and accounting data of Brazilian firms sourced from the Brazilian Securities and Exchange Commission (CVM).

Liquidity and Coverage Ratios

Column Name	Ratio Name	Formula
A36	Current Ratio	Current Assets / Current Liabilities
A37	Quick Ratio	(Current Assets - Inventory) / Current Liabilities
A38	Cash Ratio	Cash and Cash Equivalents / Current Liabilities
A44	Liquidity Ratio	Liquid Assets / Current Liabilities
A41	Tangible Asset Coverage Ratio	Tangible Assets / Total Liabilities
A43	Ratio of Commitments to Tangible Assets	Commitments / Tangible Assets

Leverage Ratios

Column Name	Ratio Name	Formula
A39	Interest Coverage Ratio	EBIT / Interest Expenses
A40	Debt Ratio	Total Liabilities / Total Assets
A42	Ratio of Equity to Debt	Total Equity / Total Debt
A48	Current Debt Ratio	Current Liabilities / Total Assets
A71	Financial Leverage	Total Debt / Total Equity
A72	Operational Leverage	Contribution Margin / Operating Profit
A73	Combined Leverage	Financial Leverage × Operational Leverage

Activity Ratios

Column Name	Ratio Name	Formula
A45	Receivable Assets Ratio	Receivables / Total Assets
A46	Fixed Asset Ratio (FAR)	Fixed Assets / Total Assets
A47	Ratio of Stockholders' Equity to Fixed Assets	Stockholders' Equity / Fixed Assets
A50	Ratio of Receivables to Gross Income	Receivables / Gross Income
A53	Turnover Ratio of Account Payable	Net Credit Purchases / Average Accounts Payable
A54	Turnover of Current Assets	Net Sales / Average Current Assets
A56	Total Capital Turnover	Net Sales / Total Capital

Figure 3.2 – Dataset – subset columns – Liquidity ratios, Leverage ratios, Activity ratios

Profitability Ratios

Column Name	Ratio Name	Formula
A49	Operating Net Profit Ratio	Operating Net Profit / Total Revenue
A57	Return On Assets (ROA)	Net Income / Total Assets
A58	Ratio of Net Profit to Total Assets	Net Profit / Total Assets
A59	Ratio of Net Profit to Current Assets	Net Profit / Current Assets
A61	Return On Equity (ROE)	Net Income / Shareholder's Equity
A62	Operating Profit Ratio	Operating Profit / Net Sales

Cost and Expense Ratios

Column Name	Ratio Name	Formula
A63	Ratio of Total Operating Cost to Gross Revenue	Total Operating Cost / Gross Revenue
A64	Expenses to Sales Ratio (ER)	Total Expenses / Net Sales
A65	Management Expense Ratio (MER)	Management Expenses / Total Revenue
A66	Financial Expense Ratio (FER)	Financial Expenses / Total Revenue

Cash Flow Ratios

Column Name	Ratio Name	Formula
A67	Free Cash Flow (FCF)	Operating Cash Flow - Capital Expenditures
A68	Ratio of Operating Cash to Net Profit	Operating Cash Flow / Net Profit
A69	Ratio of Operating Cash to Income	Operating Cash Flow / Income
A70	Cash Recovery Rate	Cash Received / Cash Invested

Per share ratios

Column Name	Ratio Name	Formula
A82	Earnings per Share	Net Income / Total Outstanding Shares
A83	Net Asset Value per Share (NAVPS)	(Total Assets - Total Liabilities) / Total Outstanding Shares
A84	Net Cash per Share	(Cash and Cash Equivalents - Total Liabilities) / Total Outstanding Shares

Figure 3.3 – Dataset – subset columns – Profitability ratios, Cost and expense ratios, Cash flow ratios, Per share ratios

The dataset given in Fig 3.2 and 3.3, spans ten years from 2011 to 2020 and consists of 23,834 records from 905 distinct corporations, each characterized by 84 indicators. Just 651 businesses encountered financial trouble, whilst the majority of businesses exhibited no financial difficulties. The data shows a significant imbalance, with 2.73% of the data pertaining to enterprises

experiencing financial hardship, and 97.27% not. This dataset is collected from Kaggle, shared by Rubens Marques Chave (Kaggle - Rubens Marques Chaves, 2023).

This dataset comprises accounting and finance data from the balance sheet, income statement, and cash flow statement, along with derived ratios and metrics. However, this study focuses on using ratios as the input, as these comprehensive ratios are derived from data available in the balance sheet, income statement, and cash flow statement. The list of ratios and their corresponding formulas are provided in Figure 3.2 and 3.3. These ratios are grouped under specific headers and subsequently processed using Principal Component Analysis (PCA) before modeling.

3.4 Tools

3.4.1 Software

- Python
- Machine learning libraries (Tensorflow, Scikit learn, Statmodels, Imblearn, Fancyimpute)
- Visualization tools (Seaborn, Matplotlib, Pacmap)
- Data processing and statistical libraries (Numpy, Pandas, scipy)
- Version control (git)
- Development environment (Vscode, Jupyter notebook)

3.4.2 Hardware

- Laptop
- Ram: 16 GB or higher
- Graphics card: Integrated GPU (Intel Iris Xe Graphics or higher)
- Processor: 10th generational Intel Core i7 or equivalent

- Storage: 256GB SSD

3.5 Summary

Studies have attempted to address these challenges using techniques like MICE and Autoencoders for missing value imputation and other methods for class imbalance. However, these approaches have not been widely explored in combination. This study aims to preserve original data patterns by integrating these methods while combining various financial ratios under respective categories to maintain their integrity in predicting financial distress. By addressing data quality and class imbalance, this research offers a more accurate and realistic approach to financial distress prediction, enhancing both predictive accuracy and real-world applicability.

CHAPTER 4: ANALYSIS

4.1 Introduction

This section provides an in-depth analysis of the insights gained through Exploratory Data Analysis (EDA) and the code-level implementation of the methodologies used. It focuses on the practical application and interpretation of the techniques, offering a detailed understanding of the processes involved.

4.2 Data Preparation

As outlined in Section 3.3.1.2, unnecessary columns and rows were removed based on the following criteria:

- All non-ratio columns were dropped.
- Rows with more than three null values were eliminated, ensuring that the overall null value threshold remained at 30% of the entire dataset.

4.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to uncover the underlying patterns and structure of the dataset, informing the selection of imputation, imbalance handling, and modeling techniques.

4.4 Distribution and Pattern Analysis

Skewness and kurtosis analyses revealed significant deviations from normal distribution, with high skewness in variables like A39, A46, A50, and A57 indicating the presence of outliers or non-normal distribution. Conversely, as observed in Fig. 4.1, variables such as A41 and A71 exhibited negative skewness, suggesting left-skewed distributions. The high kurtosis in several variables further pointed to heavy-tailed distributions and sharp peaks, likely indicating extreme values.

Variable	Skewness	Kurtosis	Pearson	Spearman	Difference
A36	11.853959	214.616444	-0.052785	-0.186587	0.133803
A37	12.563713	238.387962	-0.049422	-0.193744	0.144323
A38	19.230577	719.414609	-0.034818	-0.198795	0.163977
A39	92.826627	9554.09276	-0.006026	-0.154262	0.148236
A40	23.176508	678.464618	0.076124	0.208907	-0.132783
A41	-44.478	2189.159632	-0.075268	-0.12812	0.052852
A42	15.575324	429.057255	-0.07156	-0.208907	0.137347
A43	-53.722598	5832.43493	-0.000663	-0.058906	0.058243
A44	15.575324	429.057255	-0.07156	-0.208907	0.137347
A45	0.621959	-0.11503	0.047827	0.045094	0.002733
A46	-21.737562	493.791839	0.005761	-0.159709	0.16547
A47	99.696076	10705.21666	-0.003122	-0.107671	0.104549
A48	52.56419	3246.735664	0.060547	0.186587	-0.12604
A49	101.72517	11054.70961	0.018526	-0.003411	0.021937
A50	105.92715	11819.77462	-0.001935	0.103271	-0.105206
A53	112.645942	13037.69746	-0.002296	-0.156048	0.153752
A54	74.679641	6078.258053	-0.003347	-0.072869	0.069522
A56	25.852578	742.574246	-0.007283	-0.070814	0.063531
A57	112.588181	13046.474	-0.000546	0.098465	-0.09901
A58	109.329759	12720.48229	-0.001641	-0.15635	0.154709
A59	26.478478	5716.021951	-0.001936	-0.139285	0.137349
A61	109.457276	12499.41638	-0.001296	0.031274	-0.03257
A62	84.054508	7083.765482	-0.001792	-0.104912	0.103119
A63	101.70797	11051.89469	0.020103	0.099091	-0.078988
A64	115.887423	13611.93549	-0.000357	0.168774	-0.169131
A65	115.988407	13626.32073	-0.001759	0.151417	-0.153175
A66	39.638575	2008.157486	0.004019	0.12145	-0.117431
A67	12.941894	327.421332	-0.019246	0.001078	-0.020324
A68	84.154659	7106.654294	-0.001843	-0.081364	0.079521
A69	32.066753	1427.410195	0.006901	-0.07665	0.083552
A70	38.35699	2166.99992	-0.013655	-0.121491	0.107836
A71	-80.517737	7017.241129	0.001437	-0.065148	0.066585
A72	10.844683	2767.817807	-0.002777	-0.02006	0.017283
A73	-113.639109	13303.18985	0.001338	-0.006435	0.007773
A74	112.127483	12991.82657	-0.001139	-0.062796	0.061657
A75	42.381886	2069.384189	-0.005089	-0.033783	0.028694
A76	97.261899	10467.21012	-0.003051	-0.040991	0.03794
A77	97.681224	10531.65797	0.000136	-0.01927	0.019406
A78	79.427543	9905.807411	-0.000566	-0.007468	0.006902
A79	-74.025765	6193.009085	0.001011	-0.037715	0.038726
A80	42.900825	2084.388509	-0.008335	-0.077398	0.069062
A81	-73.467769	7896.140975	0.001957	-0.060608	0.062565
A82	44.212369	2142.866054	-0.004526	-0.139033	0.134508
A83	-41.6258	1910.806735	0.004567	-0.117571	0.122138
A84	34.122772	1263.921119	-0.005271	0.060999	-0.06627
LABEL	6.533223	40.688804	NaN	NaN	NaN

Figure 4.1 – Distribution analysis (Skewness, Kurtosis, Pearson, Spearman values) of original data

Moreover, both Pearson and Spearman correlation analyses showed relatively low correlation coefficients, indicating weak linear and monotonic relationships among most variables. However,

the difference between Pearson and Spearman correlations in most of the variables hints at potential non-linear relationships. This implies that linear models may not fully capture the underlying dynamics in the data, making non-linear models more suitable for effective prediction. While linear models alone may not deliver optimal performance, incorporating more advanced models, such as decision trees (DT), random forests (RF), gradient boosting (GBM), and artificial neural networks (ANN), proves to be particularly effective for this dataset. These models are better suited to capturing and modeling complex non-linear relationships.

4.4.1 Outliers

As mentioned in 3.2.1, Isolation forests and Z-score were employed to capture a comprehensive range of outliers within the dataset. As seen in Fig. 4.2, Z-score has identified more outliers than isolation forests due to the skewed nature of the dataset. The Z-score method's sensitivity to deviations from the mean makes it more likely to flag points as outliers in skewed and kurtotic distributions, where extreme values are more common. Unlike Z-score, the isolation forest is less sensitive to the skewness as it does not rely on measures like the mean.

Method	Outliers
Isolation forest	1402
Z-score	2048

Figure 4.2 – Outliers of Original data

The presence of untreated outliers will play a crucial role in the robustness of subsequent analyses, particularly in evaluating the effectiveness of imputation techniques and handling class imbalances. Further, the decision to leave these outliers untreated aligns with the research's focus of maintaining the dataset's originality, enabling a more accurate exploration of data preprocessing techniques in real-world scenarios.

4.4.2 Correlation Analysis

The heatmap provided in Fig. 4.3, represents the correlation matrix of variables in the dataset.

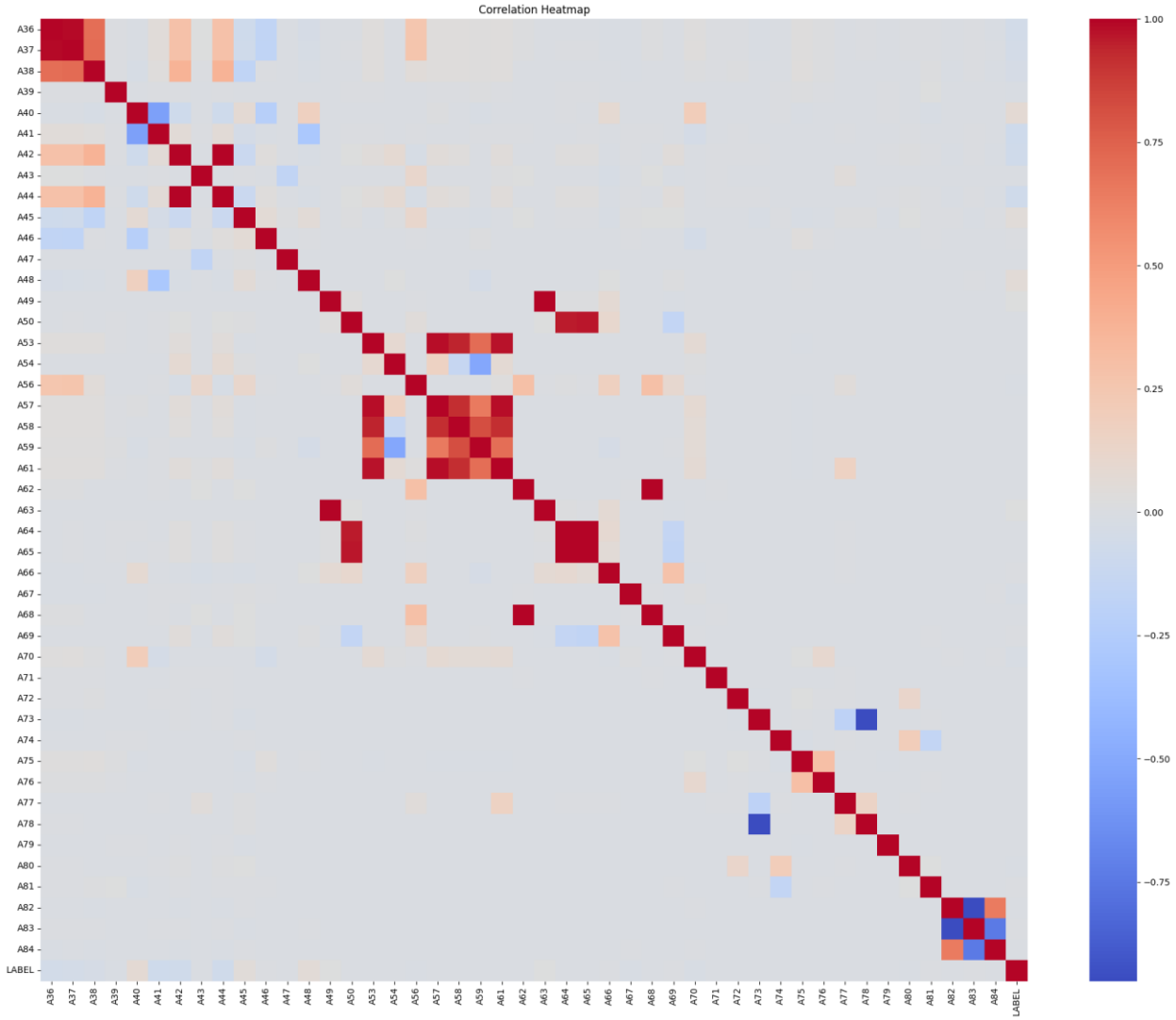

Figure 4.3 – Correlation analysis of Original data

The correlation matrix reveals several clusters of strongly correlated variables, indicating significant redundancy in the dataset. These clusters, particularly around the blocks involving variables A49 to A57 and A80 to A84, suggest that certain groups of variables are closely related. These could potentially represent similar financial ratios. These strong intra-group correlations suggest that many variables provide overlapping information, which can lead to multicollinearity

issues in modeling and leave out vital information. Further, the high-dimensional nature of the dataset, capturing vital information poses a challenge for the models. To mitigate these challenges, PCA was used to reduce redundancy by condensing correlated variables into a smaller set of components organized under grouped headers. These components were then combined to create a dataset for further processing. This approach ensures that the patterns and contributions from all the ratios are effectively captured and utilized.

4.4.3 t-SNE Graph

Unlike linear dimensionality reduction techniques such as PCA, t-SNE is capable of capturing non-linear relationships between data points, which makes it a powerful tool for uncovering complex structures in the data. It is particularly useful for visualizing complex, high-dimensional data in a lower-dimensional space, hence t-SNE has been employed.

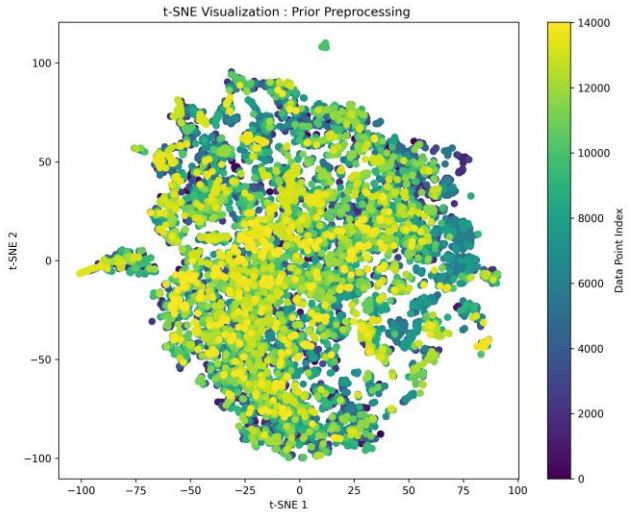

Figure 4.4 – t-SNE of original data

The plot in Fig. 4.4, reveals a densely populated central region, where data points cluster tightly together, lacking distinct boundaries or clear separations. This suggests a highly complex dataset with overlapping features and potentially intertwined classes, where the relationships between variables are intricate and non-linear. The absence of sharp boundaries further underscores the difficulty in separating these relationships in lower dimensions.

Despite the dense, interconnected appearance, there are subtle variations in point density across the plot, hinting at the presence of local structures or micro-clusters. These micro-clusters may represent subgroups of data points that are more similar to each other than to the rest of the dataset.

The smoothly distributed colour gradient, representing the data point index, indicates a continuous transition of data across the plot, rather than random scattering. This smooth gradient suggests that the data is interconnected, with gradual transitions between different subgroups or features

Given these insights, ADASYN was selected for its ability to mimic smooth transitions in the data, KMeans SMOTE (KMSMOTE) for its robustness in handling micro-clusters, and SVMSMOTE for its potential to identify and separate data points that lack sharp boundaries.

4.4.4 PaCMAP Graph

The PaCMAP (Pairwise Controlled Manifold Approximation Projection) 3D visualization provides a detailed glimpse into the high-dimensional structure of the dataset, revealing complex patterns and relationships between data points categorized as "Distressed" and "Normal".

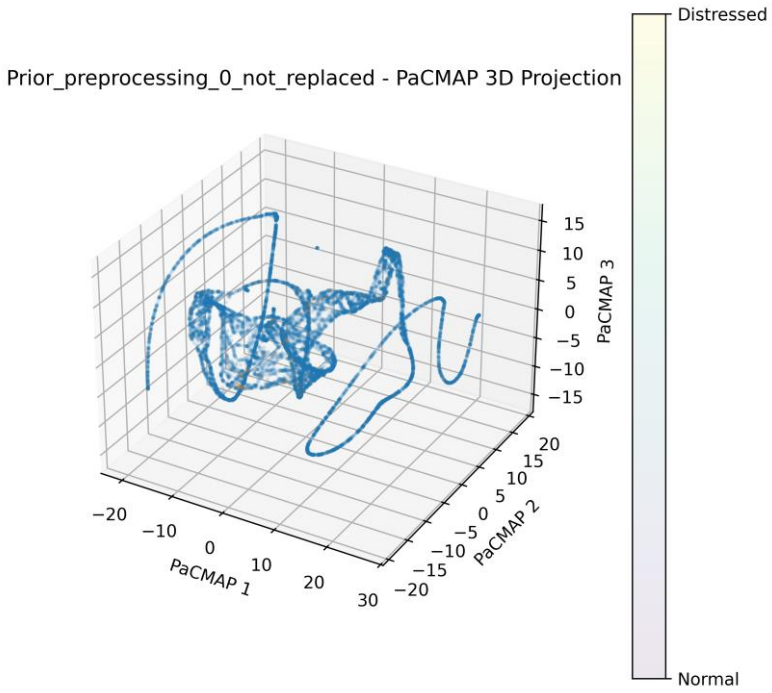

Figure 4.5 – PaCMAP analysis – Original data

This 3D projection given in Fig. 4.5, indicates that the dataset has a highly intricate structure. The data points are not merely scattered or linearly separable; instead, they form a complex, intertwined structure. This suggests that the underlying relationships between the features are non-linear and multi-faceted, making traditional linear models insufficient for capturing the full scope of the data's variance. The overlap of classes is also indicated as there is no sharp boundaries identified.

The graph shows highly imbalanced data and shows a spread-out pattern of “distressed” and “normal” firms. The distribution appears to be skewed, with the "Normal" class dominating the dataset, a common scenario in financial datasets where distressed cases are less frequent.

The transparency of the data points has been intentionally kept low to enhance the clarity of the clusters of data points. The lack of clear separation between classes, combined with the data imbalance, suggests that traditional classification models may struggle with this dataset.

This PaCMAP 3D projection reveals a dataset characterized by high complexity and overlapping class structures. This visualization highlights the challenges of modeling such data, emphasizing the need for sophisticated techniques to address the class imbalance and preserve the non-linear relationships along with the intricate patterns inherent in the dataset.

4.5 Implementation

The EDA, along with the modeling and other related processes, was implemented locally. The code was structured into distinct segments, with separate scripts dedicated to preprocessing, and model training-evaluation. After applying each Section of methodologies, the resulting datasets were stored locally to facilitate access by other scripts.

4.5.1 Preprocessing

4.5.1.1 Imputation

As discussed in section 3.2.2, two advanced imputation techniques were employed to handle missing data: Multiple Imputation by Chained Equations (MICE) and an Autoencoder-based approach. These methods were selected to ensure that the imputed datasets preserved the underlying patterns and relationships within the data, which is crucial for the accuracy and reliability of subsequent modeling.

4.5.1.1.1 MICE

The MICE method was implemented using a RandomForestRegressor as the estimator. The primary advantage of using RandomForestRegressor lies in its ability to model complex, non-linear relationships between variables, making it well-suited for datasets with intricate patterns. The RandomForestRegressor was configured with 50 estimators, which strikes a balance between

computational efficiency and the robustness of the imputation. The MICE algorithm iteratively imputed missing values by predicting them based on other available data points. The MICE algorithm was configured to perform 10 iterations, allowing the model to iteratively refine the imputed values.

4.5.1.1.1 Autoencoder

The Autoencoder-based imputation was implemented using a neural network architecture tailored to capture and reconstruct the dataset's latent structure. Initially, missing values in the dataset were replaced with zeros, a preprocessing step that allowed the neural network to process the data uniformly.

The Autoencoder architecture consisted of an input layer corresponding to the number of features, followed by a bottleneck layer (encoding layer). This bottleneck layer used the ReLU activation function, which introduced non-linearity into the model, allowing it to capture complex relationships. The output layer (decoding layer) aimed to reconstruct the original input, using a linear activation function to ensure the output was on the same scale as the input.

The model was compiled with the Adam optimizer and the mean squared error (MSE) loss function, chosen for its effectiveness in regression tasks like imputation. The Autoencoder was trained for 35 epochs with a batch size of 32, balancing training time and the model's ability to learn from the data. A validation split of 20% was used to monitor the model's performance and avoid overfitting.

After training, the Autoencoder was used to predict the missing values by reconstructing the entire dataset. The predicted values replaced the zeros in the original dataset where missing values had been.

4.5.1.2 Class Imbalance

In the analysis phase of this study, addressing class imbalance was a crucial step to ensure that the predictive models could accurately learn from and represent both the minority and majority classes.

As discussed in section 3.2.2, to achieve this, three advanced resampling techniques were implemented such as KMeans-SMOTE, SVMSMOTE, and ADASYN.

4.5.1.2.1 ADASYN

ADASYN was then applied with the intent to adaptively generate more synthetic samples in areas where the minority class was underrepresented. This method effectively focuses on the hardest-to-learn instances of the minority class, ensuring that the synthetic samples are well-distributed across the feature space. After resampling, the data was converted back to its original scale, and the class distribution was verified.

4.5.1.2.2 KMeans SMOTE (KMSMOTE)

The KMeans SMOTE process utilized a KMeans estimator with 5 clusters, a deliberate choice to better capture the structure of the minority class within the feature space. The method was then applied with a cluster balance threshold of 0.01, ensuring that the synthetic samples generated were well-distributed across the identified clusters.

4.5.1.2.3 SVMSMOTE

The implementation began with the standardization of features using StandardScaler, which normalized the data and ensured that all features contributed equally to the SVM's decision boundary. The SVMSMOTE method was then applied, which, guided by an SVM model, created synthetic samples in the feature space that are close to the minority class's decision boundary. This approach ensures that the new synthetic samples are effective in helping the model distinguish between the classes. After resampling, the dataset was inverse transformed back to the original scale and verified for balanced class distribution.

Visual inspection of all the processed datasets has been implemented through t-SNE and PaCMAP.

4.5.1.3 Dimensionality reduction

To manage the complexity and high dimensionality of the dataset, Principal Component Analysis (PCA) was employed like discussed in Section 3.2.2.3. Financial ratios were grouped into categories and PCA was applied to each group to extract two principal components. Then each category's features were standardized using StandardScaler to ensure uniformity across scales. PCA was applied to each group, reducing the dimensionality while retaining the majority of the variance. The resulting components were combined into a single DataFrame, effectively condensing the dataset.

4.5.1.4 Modeling

In this section, a thorough approach was adopted to train, test, and evaluate multiple machine learning models on the dataset. The models detailed in Section 3.2.3 are discussed regarding their implementation and performance evaluation.

4.5.1.4.1 Training

4.5.1.4.1.1 ANN

The ANN was constructed using a Sequential model with an input layer matching the number of features, followed by two hidden layers with 12 and 8 neurons, respectively, using the ReLU activation function. The output layer employed a sigmoid activation function to predict binary outcomes. The model was compiled with binary cross-entropy loss and the Adam optimizer and trained for 150 epochs with a batch size of 10.

4.5.1.4.1.2 Other models

For the other models, hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation. Each model’s parameters, such as the number of estimators and max depth for RandomForest and XGBoost, the regularization parameter for SVM and Logistic Regression, were fine-tuned to find the best-performing configuration. The best estimator from each GridSearchCV was selected for final training.

4.5.1.4.2 Testing

Once trained, the models were tested on the test dataset to generate predictions. The ANN model's predictions were thresholded at 0.5 to obtain binary outcomes, while other models used their default prediction methods.

4.5.1.4.3 Evaluation

As discussed in Section 3.2.5, the evaluation phase employed a diverse set of metrics to capture various aspects of model performance.

4.5.2 Visualizations

To enhance the visualization of the data throughout the processes and to gain a deeper understanding of the structure and implementation of techniques, t-SNE (t-Distributed Stochastic Neighbor Embedding) and PaCMAP (Pairwise Controlled Manifold Approximation Projection) were employed.

t-SNE was used to project the high-dimensional data into a 2D space. The implementation involved standardizing the dataset to ensure uniformity across all features. t-SNE was configured with a perplexity of 30 and reduced the data to two dimensions, which allowed for a clear visual representation of the data's underlying structure.

PaCMAP was employed to further explore the dataset in 3D spaces. PaCMAP is known for its ability to preserve both global and local structures in the data, making it highly effective for visualizing clusters and patterns that might not be evident in other methods.

4.6 Summary

The exploratory data analysis (EDA) reveals that the dataset exhibits a non-linear pattern with prominent outliers, indicating a complex structure. The presence of strongly correlated variables suggests the need for appropriate imputation techniques and class imbalance handling methods. Consequently, suitable methods were selected and implemented, with detailed steps outlined in this chapter. The outcomes of these implementations are further discussed in Chapter 5.

CHAPTER 5: RESULTS AND DISCUSSION

5.1 Introduction

This section presents the results of the experiments outlined in Chapter 3. The findings are evaluated to identify the most effective methods for null value imputation and class imbalance handling. This evaluation involves assessing the performance of various derived datasets using a comprehensive range of predictive models.

5.2 Imputation

As discussed in Section 4.5.1.1, Autoencoders and MICE have been implemented for imputing null values. T-SNE and PaCMAP visualizations have been generated to better interpret the impact of these methodologies.

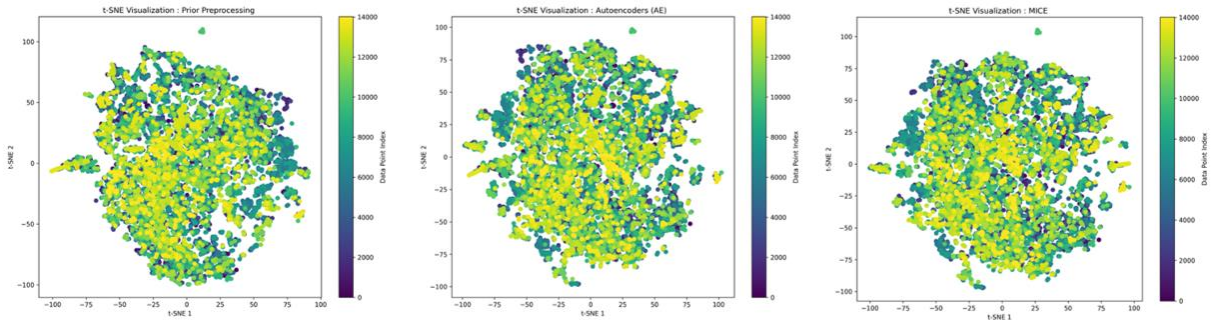

Figure 5.1 – t-SNE analysis of imputation methods

As observed from Fig. 5.1, the autoencoder imputation's t-SNE visualization reveals a more coherent and slightly more uniformly spread out points, and shows a similar structure to that of the original data. The color distribution remains consistent, indicating that the imputation process has not introduced any significant biases. The dataset's consistency improved after AE imputation, showing that it effectively filled in missing data without changing the original structure. Compared

to the raw and AE-imputed datasets, the MICE-imputed dataset exhibits stronger clustering, with data points with similar color scales more tightly grouped together. The distinct groupings indicate that MICE has successfully imputed missing values in a way that reflects the underlying dependencies in the data. While MICE has provided a clustering structure, there is also a possibility that it may have introduced some degree of overfitting by too closely adhering to the observed patterns in the non-missing data. As observed, both methods exhibit distinct strengths, Autoencoders excel in maintaining the overall structure of the data, offering a more balanced view, while MICE proves more effective in preserving and reinforcing relationships and underlying patterns, even those not immediately apparent in visualizations.

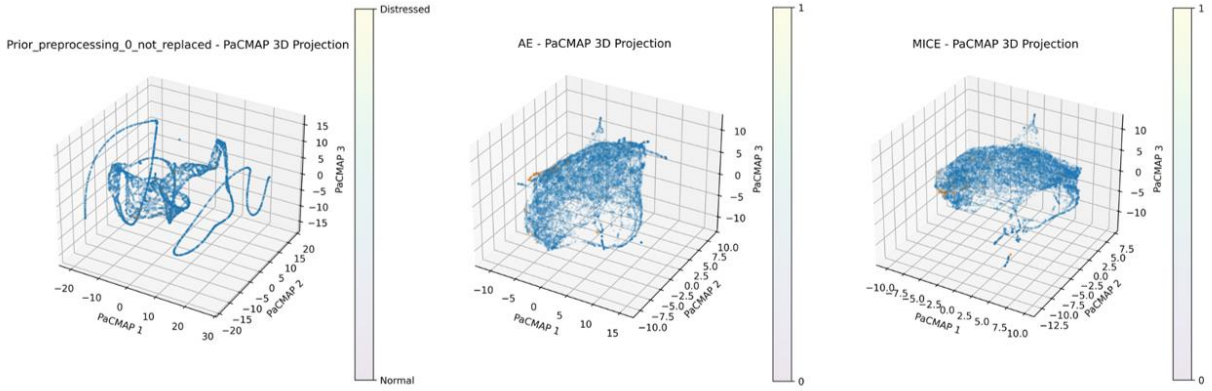

Figure 5.2 – PaCMAP analysis of imputation methods

The PaCAMP visualization of the raw data reveals a highly complex and non-linear structure as observed in Fig. 5.2. As discussed in the Section 4.3, the color transparency have been purposefully kept low for better clarity of the structure. The data points are observed to be spread out in a convoluted manner. The structure indicates the dataset is noisy with overlapping regions of classes. It also reflects the possibility of missing data, disrupting the structure of the dataset.

As previously highlighted, Autoencoders excel at capturing non-linear relationships. The more compact structure observed in the Autoencoder visualization suggests that this imputation method preserves the integrity of the data's underlying patterns while effectively filling in missing values. This approach strikes a balance between maintaining relationships and minimizing noise and variance from missing data. In contrast, MICE is particularly adept at predicting missing values by leveraging the relationships between variables. The MICE-imputed dataset displays a more tightly clustered structure, indicating that MICE imputation has strengthened the relationships between variables by rigorously filling in missing data based on observed patterns.

While both, Autoencoders and MICE have been observed to have significantly altered the structure of the original data with disruptions, Autoencoder imputation smooths out this complexity, creating a more balanced and interpretable dataset. MICE further organizes the data, resulting in a dense and consistent structure, though potentially at the cost of some variability. It is also important to note that the missing values in the raw data have significantly disrupted the original structure of the dataset.

5.3 Class Imbalance

As mentioned in Chapter 3, ADASYN, KMeans SMOTE (KMSMOTE), and SVMSMOTE have been implemented to address class imbalance issues.

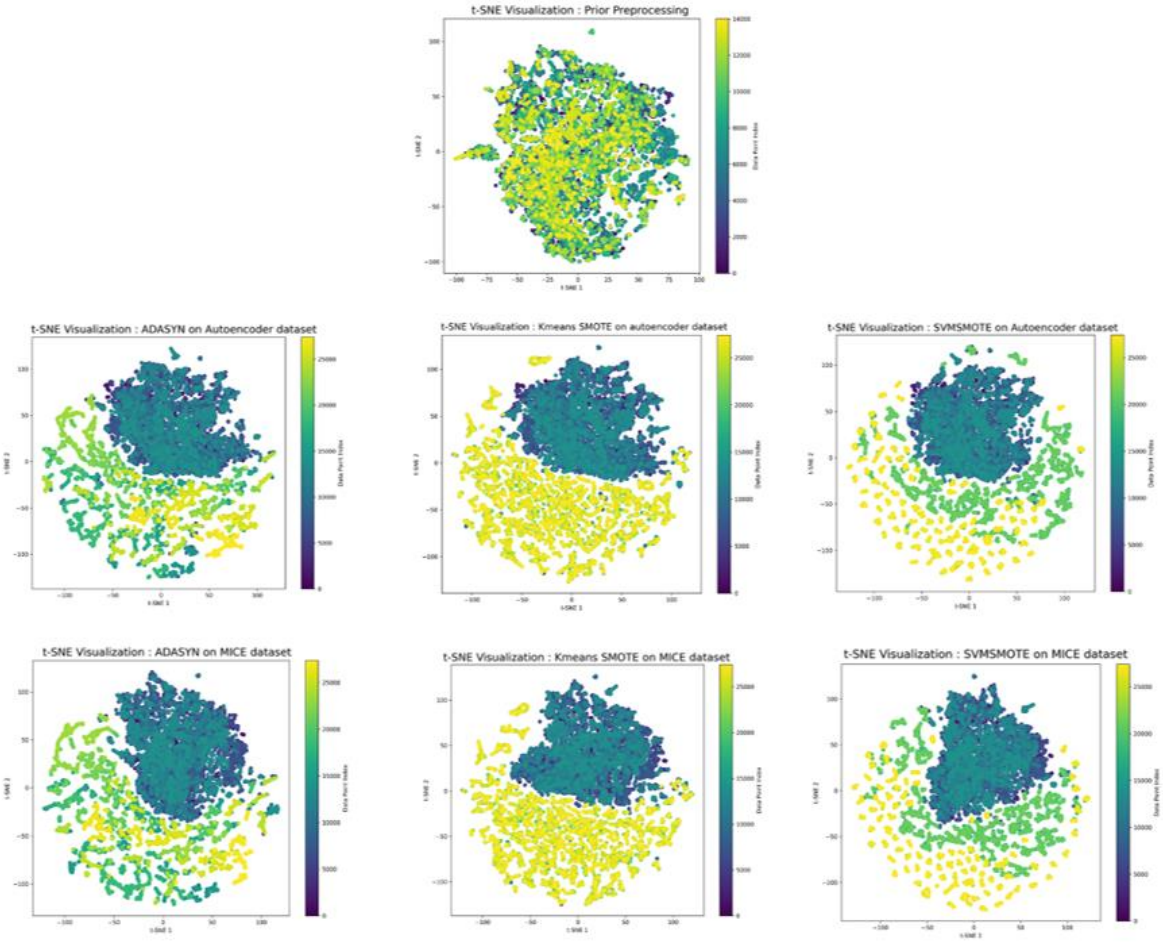

Figure 5.3 – t-SNE analysis of class imbalance methods

As detailed earlier in the sections 4.5.1.2, the raw data prior to preprocessing exhibited data point overlap and potential class imbalance. Although the differences between the data structures of Autoencoder and MICE imputation methods appear subtle across various imbalance handling techniques, each imputation method has influenced the data structure distinctively. Specifically, the ADASYN method shows a gradual separation of scales of data still having a few overlaps. It

has generated synthetic samples that mimic the natural transitions within the data, as evidenced by the smooth spread of points across the visualization.

The t-SNE plots for KMSMOTE show well-defined central clusters surrounded by more dispersed points. This proves that KMSMOTE is particularly effective at reinforcing micro-clusters within the data. However, the rigid separation of clusters could potentially limit the flexibility of the dataset in capturing more fluid relationships.

The SVMSMOTE graphs reveal a well-balanced distribution of data points, featuring a distinct central cluster surrounded by well-dispersed points. This method is effective at identifying data points with unclear boundaries and creating synthetic samples that highlight these distinctions, thus improving the overall dataset structure. As discussed earlier in the EDA, SVMSMOTE effectively clarifies the dataset by separating points in regions with unclear boundaries. To further investigate the data structures, PaCMAP graphs have also been generated.

As observed in Fig. 5.4, the raw data shows a highly convoluted and tangled structure, with overlapping classes and a heavy imbalance of data.

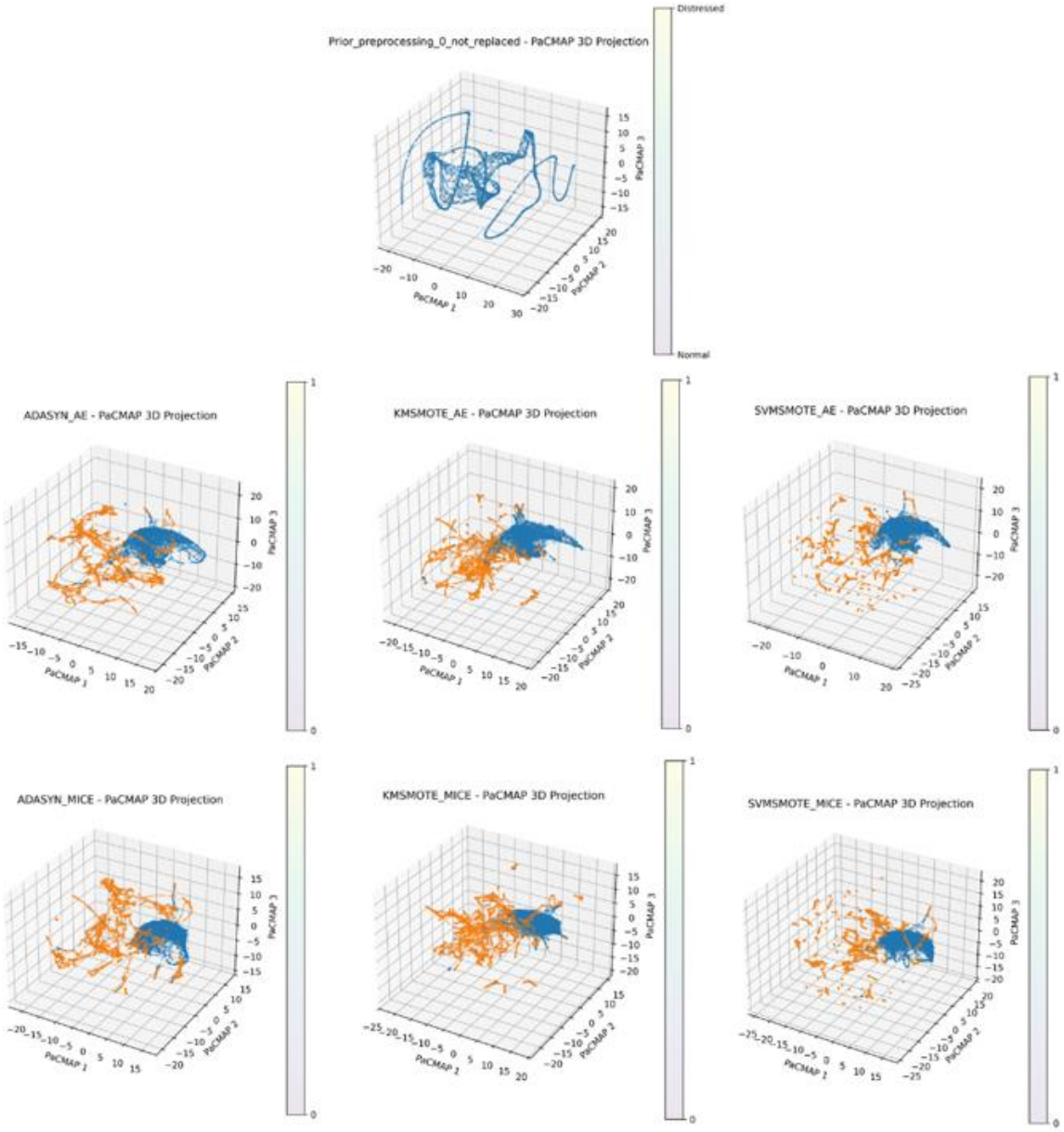

Figure 5.4 – PaCMAP analysis of class imbalance methods

The visualizations for ADASYN on both AE and MICE imputed datasets show a more dispersed and less clustered structure compared to the prior preprocessing state. This method also resembles similar patterns of raw data, having intertwined structures, with minimal classes intermixed. The spread of points suggests that ADASYN has effectively filled in the gaps within the minority class,

particularly in areas where the distinction between classes was less pronounced. The visualizations indicate that the classes are still not entirely separable, reflecting the method's focus on smooth transitions rather than sharp boundaries.

Where as the KMeans SMOTE visualizations show a more compact and clustered structure. The synthetic samples generated appear to reinforce existing clusters with tighter represented micro-clusters. This method strengthens the relationships within clusters, making the dataset more structured. However as highlighted before, it may reduce the overall variability in the dataset by concentrating synthetic samples within specific regions.

Alternatively, the SVMSMOTE visualizations reveal a balanced and dispersed structure, with data points more evenly distributed in the space. There is a noticeable separation between the classes, though the boundaries are not overly sharp, reflecting the method's ability to manage ambiguous class boundaries. These graphs prove that SVMSMOTE is particularly effective at identifying and separating data points that lack clear boundaries, as analysed.

ADASYN is observed to closely mirror the gradual transitions between classes and maintain the original convoluted structure of the dataset. It effectively captures subtle variations and gradual shifts, thereby preserving the inherent patterns present in the data. In comparison, KMeans SMOTE highlights more intricate patterns within the data, with the generation of tighter micro-clusters that reflect the complex, intertwined nature of the dataset. This method is adept at focusing on local data structures and enhancing their representation. On the other hand, SVMSMOTE is noted for its ability to distinctly separate classes with sparse and dispersed data points while accommodating the dataset's complexity. It ensures that the original patterns are preserved and significantly improves class separation.

5.4 Outliers

The outliers introduced by the class imbalance methods are analyzed to gain a deeper understanding of their impact on data distribution. This analysis helps in assessing how these methods affect the overall data structure and the representation of different classes.

Prior preprocessing

Method	Outliers
Isolation forest	1402
Z-score	2048

ADASYN AE

Method	Outliers
Isolation forest	2741
Z-score	2233

SVMSMOTE AE

Method	Outliers
Isolation forest	2742
Z-score	2014

KMSMOTE AE

Method	Outliers
Isolation forest	2738
Z-score	2000

ADASYN MICE

Method	Outliers
Isolation forest	2738
Z-score	2245

SVMSMOTE MICE

Method	Outliers
Isolation forest	2742
Z-score	1998

KMSMOTE MICE

Method	Outliers
Isolation forest	2742
Z-score	1945

Figure 5.5 – Outliers of imputed and class imbalance handled datasets

The Fig. 5.5 shows that before applying any methods, the original data reveals a significant number of outliers, comprising over 12% of the entire dataset. The Z-score method detects more outliers compared to the Isolation Forest, likely due to the dataset's skewed structure, as discussed earlier. After implementing class imbalance techniques, new outliers are introduced, but the proportion of outliers decreases to an average of 8% across the dataset. Despite this reduction, there are minimal variations in the number of outliers among the different imbalance handling methods.

The number of outliers detected by the Z-score method decreases slightly in all the derived datasets compared to the prior preprocessing stage, indicating that the method has smoothed some of the extreme values, leading to fewer detections. These findings also suggest the imputation methods have slightly altered the skewness of the original data structure. KMeans SMOTE (KMSMOTE) reduced the number of extreme values considered as outliers by the Z-score method by reinforcing micro-clusters, potentially leading to a more homogeneously distributed dataset. SVMSMOTE results in a similar number of outliers detected by the Isolation Forest as seen with KMeans SMOTE.

While ADASYN tends to increase the number of potential outliers detected by the Isolation Forest, KMeans SMOTE (KMSMOTE) and SVMSMOTE appear to manage outliers more effectively, particularly in the MICE-imputed datasets.

5.5 Model Evaluation

A comprehensive list of models, as outlined in Section 3.2.3, has been implemented, and their results are discussed in this section. To provide a clearer understanding of the processes, the models were trained on datasets that were handled with imputation methods alone, as well as on datasets that were processed with both imputation and class imbalance techniques separately.

5.5.1 Imputed Datasets

The datasets with null values imputed by MICE and Autoencoders are evaluated. The results are shown in Fig. 5.6.

Dataset	Models	Accuracy	AUC ROC	F1 Score	Precision	Recall	Kappa	G Mean
AE	ANN	0.980	0.977	0.455	0.575	0.377	0.446	0.612
AE	GradientBoosting	0.991	0.991	0.760	0.974	0.623	0.756	0.789
AE	KNN	0.985	0.919	0.568	0.794	0.443	0.562	0.664
AE	LogisticRegression	0.975	0.928	0.000	0.000	0.000	-0.005	0.000
AE	NaiveBayes	0.173	0.789	0.047	0.024	0.934	0.005	0.382
AE	RandomForest	0.989	0.990	0.644	1.000	0.475	0.639	0.690
AE	SVM	0.980	0.893	0.197	0.700	0.115	0.192	0.339
AE	XGBoost	0.991	0.995	0.762	0.909	0.656	0.758	0.809
AE	**Average**	**0.883**	**0.935**	**0.429**	**0.622**	**0.453**	**0.419**	**0.536**
MICE	ANN	0.984	0.969	0.532	0.758	0.410	0.525	0.639
MICE	GradientBoosting	0.990	0.991	0.726	0.902	0.607	0.721	0.778
MICE	KNN	0.985	0.920	0.568	0.794	0.443	0.562	0.664
MICE	LogisticRegression	0.975	0.929	0.000	0.000	0.000	-0.005	0.000
MICE	NaiveBayes	0.223	0.784	0.051	0.026	0.967	0.009	0.447
MICE	RandomForest	0.990	0.989	0.695	0.971	0.541	0.690	0.735
MICE	SVM	0.980	0.895	0.222	0.727	0.131	0.217	0.362
MICE	XGBoost	0.989	0.994	0.700	0.897	0.574	0.695	0.757
MICE	**Average**	**0.890**	**0.934**	**0.437**	**0.634**	**0.459**	**0.427**	**0.548**

Figure 5.6 – Performance evaluation of imputed datasets

As observed, on average, both AE and MICE datasets demonstrate strong performance in terms of Accuracy with 0.88 for AE and 0.89 for MICE, and AUC with 0.93 for both the variations respectively. The MICE dataset shows a marginal improvement in both F1 Score (MICE: 0.437, AE: 0.429) and G-Mean (MICE: 0.548, AE: 0.536), suggesting that MICE imputation better balances sensitivity and specificity. However, the overall low F1 scores across models highlight that despite high accuracy, the models struggle to balance precision and recall, indicating that class imbalance affects performance. ANN, gradient boosting and XGboost are observed to perform well across both datasets with imbalance issues. XGBoost achieves the highest AUC ROC scores (AE: 0.995, MICE: 0.994) and strong F1 Scores (AE: 0.762, MICE: 0.700), indicating it is highly effective at distinguishing between classes and maintaining a good balance between precision and recall. Random Forest performs particularly well with MICE imputation, achieving one of the highest G-Mean scores (0.735), indicating robust performance across true positive and true negative rates. Logistic Regression and Naive Bayes perform poorly across both imputation methods, with particularly low F1 scores and G-Means. Ensemble methods like Gradient Boosting and XGBoost outperform other models, likely due to their ability to handle non-linear relationships and interactions within the data.

Despite strong performance in terms of accuracy and AUC ROC, the generally low F1 scores and G-Means indicate ongoing challenges with class imbalance. The low scores indicate that the models struggle to achieve a balance between correctly identifying minority class instances (recall) and minimizing false positives (precision). The low kappa statistic suggests that while the models are achieving high accuracy, their predictions do not consistently align with actual outcomes across all classes, reflecting a potential bias towards the majority class. G-Mean (AE: 0.536, MICE: 0.548) provides a balanced measure of a model's performance by considering both sensitivity (recall) and specificity. The relatively low G-Mean values indicate that the models are not performing consistently across both classes, particularly struggling with the minority class.

Dataset	Type 1 Error Rate (average)	Type 2 Error Rate (average)
AE	0.547	0.107
MICE	0.541	0.101

Figure 5.7 – Type 1 and Type 2 error rates of imputed datasets

The Fig. 5.7 provided show the Type 1 (false positive) and Type 2 (false negative) error rates for datasets imputed using AE and MICE. These error rates are critical indicators of how well the models perform, particularly in the presence of an imbalanced dataset.

The Type 1 error rates are notably high, indicating that the models frequently misclassify non-distressed companies as distressed. This suggests that the models are struggling to correctly identify the majority class, which may result from overfitting to the minority class. The Type 2 error rates, though lower than the Type 1 errors, are still concerning as this is achieved at the expense of a high Type 1 error rate. A Type 2 error indicates that the model fails to identify distressed companies, which could have severe implications in financial contexts. The high Type 1 error rates and moderate Type 2 error rates reflect the challenges posed by an imbalanced dataset, where the majority class (non-distressed companies) vastly outnumbers the minority class. The models seem to favor identifying the minority class (distressed companies), which results in lower Type 2 error rates but at the cost of a significant increase in Type 1 errors.

The results and the error classification suggest that while AE and MICE imputation methods provide a solid foundation for model training, addressing class imbalance could further enhance model performance, particularly in improving recall, F1 scores, Kappa statistic, and G-Means.

5.5.2 Combined Datasets

The datasets with null values imputed by MICE and Autoencoders were subsequently combined with class imbalance handling methods, including ADASYN, KMeans SMOTE, and SVMSMOTE. The results of these combined approaches were evaluated and are presented in Fig. 5.8.

Dataset	Models	Accuracy	AUC ROC	F1 Score	Precision	Recall	Kappa	G Mean
ADASYN AE	ANN	0.979	0.992	0.979	0.963	0.995	0.957	0.978
ADASYN AE	GradientBoosting	0.991	1.000	0.991	0.982	1.000	0.982	0.991
ADASYN AE	KNN	0.985	0.991	0.985	0.970	1.000	0.969	0.984
ADASYN AE	LogisticRegression	0.866	0.905	0.872	0.834	0.913	0.732	0.865
ADASYN AE	NaiveBayes	0.536	0.709	0.675	0.519	0.963	0.073	0.325
ADASYN AE	RandomForest	0.992	1.000	0.992	0.984	0.999	0.983	0.992
ADASYN AE	SVM	0.913	0.951	0.915	0.891	0.940	0.825	0.912
ADASYN AE	XGBoost	0.993	1.000	0.993	0.986	1.000	0.986	0.993
ADASYN AE	**Average**	**0.907**	**0.943**	**0.925**	**0.891**	**0.976**	**0.813**	**0.880**
ADASYN MICE	ANN	0.977	0.991	0.978	0.960	0.996	0.954	0.977
ADASYN MICE	GradientBoosting	0.993	1.000	0.993	0.986	1.000	0.985	0.993
ADASYN MICE	KNN	0.985	0.992	0.985	0.971	1.000	0.970	0.985
ADASYN MICE	LogisticRegression	0.875	0.914	0.880	0.840	0.925	0.749	0.873
ADASYN MICE	NaiveBayes	0.579	0.733	0.695	0.544	0.963	0.158	0.434
ADASYN MICE	RandomForest	0.992	1.000	0.992	0.984	1.000	0.984	0.992
ADASYN MICE	SVM	0.921	0.957	0.924	0.895	0.955	0.843	0.921
ADASYN MICE	XGBoost	0.993	1.000	0.993	0.986	1.000	0.985	0.993
ADASYN MICE	**Average**	**0.914**	**0.948**	**0.930**	**0.896**	**0.980**	**0.829**	**0.896**
KMSMOTE AE	ANN	0.976	0.992	0.977	0.957	0.997	0.952	0.976
KMSMOTE AE	GradientBoosting	0.991	1.000	0.991	0.984	0.999	0.983	0.991
KMSMOTE AE	KNN	0.982	0.990	0.982	0.967	0.998	0.964	0.982
KMSMOTE AE	LogisticRegression	0.910	0.948	0.911	0.901	0.922	0.820	0.910
KMSMOTE AE	NaiveBayes	0.557	0.818	0.687	0.531	0.972	0.113	0.370
KMSMOTE AE	RandomForest	0.993	1.000	0.993	0.986	0.999	0.985	0.993
KMSMOTE AE	SVM	0.938	0.977	0.938	0.926	0.951	0.875	0.937
KMSMOTE AE	XGBoost	0.991	1.000	0.991	0.984	0.999	0.983	0.991
KMSMOTE AE	**Average**	**0.917**	**0.966**	**0.934**	**0.904**	**0.980**	**0.834**	**0.894**
KMSMOTE MICE	ANN	0.984	0.995	0.985	0.974	0.995	0.969	0.984
KMSMOTE MICE	GradientBoosting	0.994	1.000	0.994	0.989	0.999	0.988	0.994
KMSMOTE MICE	KNN	0.981	0.990	0.981	0.965	0.997	0.961	0.981
KMSMOTE MICE	LogisticRegression	0.915	0.952	0.916	0.905	0.926	0.829	0.914
KMSMOTE MICE	NaiveBayes	0.591	0.844	0.704	0.552	0.971	0.181	0.452
KMSMOTE MICE	RandomForest	0.993	1.000	0.993	0.988	0.999	0.987	0.993
KMSMOTE MICE	SVM	0.941	0.978	0.942	0.926	0.959	0.883	0.941
KMSMOTE MICE	XGBoost	0.993	1.000	0.994	0.988	0.999	0.987	0.993
KMSMOTE MICE	**Average**	**0.924**	**0.970**	**0.938**	**0.911**	**0.981**	**0.848**	**0.907**
SVMSMOTE AE	ANN	0.980	0.994	0.980	0.963	0.998	0.960	0.980
SVMSMOTE AE	GradientBoosting	0.993	1.000	0.993	0.987	1.000	0.987	0.993
SVMSMOTE AE	KNN	0.990	0.994	0.990	0.981	1.000	0.980	0.990
SVMSMOTE AE	LogisticRegression	0.904	0.954	0.904	0.910	0.897	0.809	0.904
SVMSMOTE AE	NaiveBayes	0.534	0.761	0.672	0.519	0.954	0.068	0.330
SVMSMOTE AE	RandomForest	0.994	1.000	0.994	0.990	0.999	0.989	0.994
SVMSMOTE AE	SVM	0.939	0.977	0.940	0.926	0.954	0.877	0.938
SVMSMOTE AE	XGBoost	0.995	1.000	0.995	0.991	0.999	0.990	0.995
SVMSMOTE AE	**Average**	**0.916**	**0.960**	**0.934**	**0.908**	**0.975**	**0.832**	**0.891**
SVMSMOTE MICE	ANN	0.981	0.995	0.982	0.966	0.998	0.962	0.981
SVMSMOTE MICE	GradientBoosting	0.993	1.000	0.994	0.987	1.000	0.987	0.993
SVMSMOTE MICE	KNN	0.990	0.995	0.990	0.981	0.999	0.980	0.990
SVMSMOTE MICE	LogisticRegression	0.913	0.956	0.913	0.907	0.920	0.825	0.913
SVMSMOTE MICE	NaiveBayes	0.563	0.783	0.687	0.535	0.958	0.127	0.401
SVMSMOTE MICE	RandomForest	0.994	1.000	0.994	0.990	0.999	0.988	0.994
SVMSMOTE MICE	SVM	0.942	0.979	0.943	0.928	0.960	0.885	0.942
SVMSMOTE MICE	XGBoost	0.994	1.000	0.994	0.989	1.000	0.988	0.994
SVMSMOTE MICE	**Average**	**0.921**	**0.963**	**0.937**	**0.910**	**0.979**	**0.843**	**0.901**

Figure 5.8 – Performance evaluation of imputed and class imbalance handled datasets

Across all resampling methods, there is a noticeable improvement in the average performance metrics compared to the datasets that were only imputed but not balanced. The average F1 Scores, G-Means, and Kappa values have increased significantly, indicating that the models are now better at balancing precision and recall, leading to more reliable predictions. The AUC ROC scores remain consistently high across all methods, indicating strong discriminative power of the models in distinguishing between the classes. This suggests that resampling methods like ADASYN, KMSMOTE, and SVMSMOTE have effectively enhanced the models' ability to identify both classes, reducing the bias towards the majority class. While higher accuracy and AUC were observed, the focus is placed on metrics such as F1, G-mean, and Kappa which more effectively highlighted the challenges associated with the imbalanced dataset.

The average F1 Score (0.925) and G-Mean (0.880) for the ADASYN AE dataset indicate strong performance, particularly with Gradient Boosting and XGBoost, both of which achieve near-perfect AUC ROC scores and high F1 Scores. However, the performance is slightly lower compared to KMeans SMOTE and SVMSMOTE, suggesting that while ADASYN effectively balances the classes, it may not handle the most complex relationships as robustly as the other methods. The average Kappa of 0.813 suggests that while ADASYN, combined with AE, improves class balance, it might not handle the complexities of data structure as effectively as KMeans SMOTE or SVMSMOTE. The performance metrics of ADASYN MICE have been observed to be slightly higher in F1 Score (0.930) and G-mean (0.896) compared to the AE variation.

KMSMOTE AE datasets show robust performance across all metrics, with an average F1 Score of 0.934 and G-Mean of 0.894. The models perform consistently well, particularly with ensemble methods like XGBoost and Gradient Boosting. The Kappa value of 0.834 for KMSMOTE AE is among the highest, indicating strong agreement between predicted and actual outcomes. This indicates that KMeans SMOTE is particularly effective in creating well-defined clusters that models can learn from efficiently. The KMeans SMOTE MICE dataset shows the highest performance among the resampling methods, with an average F1 Score of 0.938 and G-mean of 0.907. This indicates that KMSMOTE, when combined with MICE, creates a highly balanced dataset that allows models to perform optimally across all metrics.

SVMSMOTE AE also demonstrates strong performance, particularly with XGBoost, which achieves perfect scores in several metrics, including AUC ROC, F1 Score, and G-Mean. The average performance across models is very high, with an F1 Score of 0.934, G-Mean of 0.890, and kappa of 0.832, indicating that SVMSMOTE effectively separates data points with ambiguous boundaries, leading to more accurate classifications. The SVMSMOTE MICE dataset has been observed to perform exceptionally well, with an average F1 Score of 0.937 and G-Mean of 0.901. The combination of SVMSMOTE and MICE results in models that are both highly sensitive and specific, as reflected in the high recall and G-Mean values.

On overall, most models, particularly ensemble methods like Gradient Boosting and XGBoost, achieve exceptionally high accuracy scores, often above 0.98. Ensemble methods like Gradient Boosting and XGBoost have been observed to consistently outperform other models across all datasets and resampling techniques, due to their non-linear data handling. XGBoost consistently delivers the highest accuracy, averaging over 0.99 across all datasets. Additionally, it achieves the top AUC-ROC scores across all datasets and resampling techniques, frequently reaching a perfect score of 1.0 in AUC-ROC. This reflects XGBoost's strength in handling complex datasets. Despite the improvements brought by resampling, Naive Bayes and Logistic Regression continue to struggle, particularly in terms of F1 Score and G-Mean. This suggests that these models may not be well suited for handling non-linear and complex structured datasets. SVM and KNN show improved performance post-resampling but still lag behind ensemble methods. These models are more sensitive to the specific characteristics of the dataset, and while resampling helps, they may require further tuning or alternative approaches to achieve top-tier performance.

Dataset	Type 1 Error Rate (average)	Type 2 Error Rate (average)
ADASYN AE	0.024	0.162
ADASYN MICE	0.020	0.151
KMSMOTE AE	0.025	0.143
KMSMOTE MICE	0.020	0.133
SVMSMOTE AE	0.025	0.142
SVMSMOTE MICE	0.021	0.135

Figure 5.9 – Type 1 and Type 2 error rates of imputed and class imbalance handled datasets

As observed in Fig. 5.9, the Type 1 error rates across all methods and datasets are relatively low, ranging from 0.020 to 0.025. This indicates that the models are effective at minimizing false positives, demonstrating a strong ability to correctly classify non-distressed companies as non-distressed after the application of resampling techniques. However, Type 2 error rates are consistently higher, ranging from 0.133 to 0.162, reflecting the inherent difficulty in identifying distressed companies, which represent the minority class in a complex structured dataset.

The low Type 1 error rates suggest that the models are proficient at preserving the integrity of the majority class. However, reducing Type 2 errors poses a more significant challenge. The results indicate that KMeans SMOTE (KMSMOTE) and SVMSMOTE, particularly when combined with MICE, are more successful in achieving a balance between minimizing both types of errors, thereby improving the overall model performance.

The findings indicate that advanced resampling techniques like KMeans SMOTE and SVM SMOTE, especially when used alongside MICE, achieve an optimal balance between precision and recall. This is reflected in higher kappa values, and improved G-mean and F1 scores, underscoring the effectiveness of these combinations. MICE consistently enhances the performance of resampling techniques, likely due to its ability to capture subtle relationships within the data more effectively. While Autoencoders excel in preserving the overall structure, they show a slight decrease in performance compared to MICE. It is also important to note that MICE is computationally intensive and time-consuming.

Model	Average Training Time (min)
MICE	466.88
AE	0.17
ADASYN	0.01
Kmeans SMOTE	0.01
SVMSMOTE	0.02

Figure 5.10 – Time taken for all the methods

As shown in Fig. 5.10, MICE is computationally intensive and time-consuming, with significantly longer training times due to its iterative nature. MICE imputes missing data by modeling each feature with missing values as a function of other features, making it highly effective but resource-intensive. This makes MICE more suitable for applications where accuracy is paramount, and sufficient computational resources are available. Conversely, Autoencoders offer a faster alternative, delivering imputation results that are nearly comparable to those of MICE, with considerably less computational overhead.

SVMSMOTE takes slightly longer to compute (0.02 minutes more) compared to KMeans SMOTE and ADASYN due to the additional complexity involved in identifying support vectors that lie on the class boundaries for generating synthetic samples. This process adds a layer of computational effort, contributing to the marginally increased processing time.

5.6 Summary

After a comprehensive analysis of the implemented methods across various metrics, it was observed that while both imputation methods and all class imbalance handling techniques performed effectively, MICE imputation excelled in preserving relationships and subtle underlying patterns. This method showed strong compatibility with cluster-based sampling techniques such as KMeans SMOTE (KMSMOTE) and SVMSMOTE, which demonstrated a slight advantage over ADASYN. Key findings and final conclusions are further elaborated in Chapter 6.

CHAPTER 6: CONCLUSION & RECOMMENDATIONS

6.1 Discussion and Conclusion

This thesis aimed to improve the predictive accuracy of financial distress models by tackling critical challenges in financial data, such as missing values and class imbalance. The study emphasized maintaining the originality of the data through the use of advanced imputation methods, namely Multiple Imputation by Chained Equations (MICE) and Autoencoders. In conjunction with these, sophisticated class imbalance handling techniques, including ADASYN, KMeans SMOTE (KMSMOTE), and SVMSMOTE, were employed. This comprehensive approach enabled a detailed analysis of the impact of these techniques on the performance of various machine learning models.

The findings as discussed in Chapter 5 of this research highlight that MICE and Autoencoder imputation methods significantly improve the predictive performance of models by effectively handling missing data and preserving underlying data patterns. MICE, leveraging Random Forest Regressors, excelled in capturing intricate relationships within the data, which translated into slightly better performance metrics when compared to Autoencoders. However, it was also observed that MICE is more computationally intensive, making Autoencoders a viable alternative in scenarios where computational resources are limited.

In terms of class imbalance, the study found that KMeans SMOTE and SVMSMOTE were particularly effective in generating synthetic samples that improved the models' ability to distinguish between distressed and non-distressed companies. KMeans SMOTE was noted for its ability to reinforce micro-clusters within the data, leading to higher performance metrics across various models. On the other hand, SVMSMOTE was effective in managing ambiguous class boundaries, resulting in a more balanced distribution of data points and improved model accuracy. These techniques have effectively reduced the Type 1 error rate (false positives), which had previously been a significant concern before the implementation of class imbalance handling methods.

XGBoost, Random forests, and Gradient Boosting consistently outperformed other models across various metrics. XGBoost, when applied to the dataset processed with SVMSMOTE and AE, achieved the highest performance with an overall accuracy of 0.995, with an average accuracy of 0.993 across all datasets.

However, a key challenge identified in the study is the comparatively high Type 2 error rates (false negatives), which is concerning as it involves the misclassification of distressed companies as non-distressed. Although the implemented techniques have successfully reduced Type 1 errors (false positives) and maintained strong G-Mean and F1 scores, the persistence of higher Type 2 errors suggests that the models might be underestimating risk, which could have significant implications in financial distress prediction.

Despite the improvements observed with the combined imputation and resampling techniques, the study also identified ongoing challenges, comparatively high Type 2 error rates (false negatives) as given in Fig. 5.9, which is concerning as it involves the misclassification of distressed companies as non-distressed. Although the G-Mean and F1 scores across models were consistently high, indicating a strong balance between precision and recall, the elevated Type 2 error rates suggest that the models, despite their overall robustness, still face challenges in fully capturing the complexity of financial distress.

6.2 Contributions

This thesis adds the following contributions to the existing literature on financial distress prediction:

- Demonstrated the effectiveness of MICE and Autoencoders in handling missing values, leading to improved model accuracy.
- KMeans SMOTE and SVMSMOTE have been identified as effective techniques for reducing Type 1 errors, thereby enhancing overall predictive performance

- Identified the ongoing challenge of high Type 2 errors, emphasizing the need for further refinement in financial distress models.
- Provided a foundation for integrating advanced machine learning models with sophisticated data preprocessing frameworks in financial distress prediction.

6.3 Future Recommendations

Based on the findings of this research, several recommendations can be made for future studies and practical applications.

The dataset used in this study was specific to Brazilian firms. Future research should consider applying these methods to datasets from different regions and industries to validate the generalizability of the findings and identify potential regional or sector-specific nuances. Given the high Type 2 error rates, future research should prioritize strategies aimed at minimizing false negatives. This could include exploring cost-sensitive learning techniques, adjusting decision thresholds, or developing new methods that specifically target the reduction of Type 2 errors, ensuring a more balanced prediction output. While MICE has demonstrated strong effectiveness, its high computational demands present a challenge. Future research could focus on developing hybrid models that strike a balance between high accuracy and lower computational requirements, making these methods more feasible for large-scale applications.

REFERENCES

Adisa, J.A., Ojo, S.O., Owolawi, P.A. and Pretorius, A.B., (2019) Financial Distress Prediction: Principle Component Analysis and Artificial Neural Networks. In: *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*. [online] IEEE, pp.1–6. Available at: https://ieeexplore.ieee.org/document/9015884/.

Ahmed, S., Hassan, S.U., Aljohani, N.R. and Nawaz, R., (2020) FLF-LSTM: A novel prediction system using Forex Loss Function. *Applied Soft Computing Journal*, 97.

Alam, T.M., Shaukat, K., Mushtaq, M., Ali, Y., Khushi, M., Luo, S. and Wahab, A., (2021) Corporate Bankruptcy Prediction: An Approach towards Better Corporate World. *Computer Journal*, 6411, pp.1731–1746.

Al Ali, A.I., S, S.R. and Khedr, A.M., (2024) Enhancing Financial Distress Prediction through Integrated Chinese Whisper Clustering and Federated Learning. *Journal of Open Innovation: Technology, Market, and Complexity*, [online] p.100344. Available at: https://linkinghub.elsevier.com/retrieve/pii/S2199853124001380.

Antulov-Fantulin, N., Lagravinese, R. and Resce, G., (2021) Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior and Organization*, 183, pp.681–699.

Aydin, N., Sahin, N., Deveci, M. and Pamucar, D., (2022) Prediction of financial distress of companies with artificial neural networks and decision trees models. *Machine Learning with Applications*, 10, p.100432.

Balachander, T., Akhlaq, N., Bansal, R., Vasani, S.A., Singh, K. and Mannar, B.R., (2023) Financial Crisis Prediction using Feature Subset Selection with Quantum Deep Neural Network. In: *Proceedings of the 2023 2nd International Conference on Electronics and Renewable Systems, ICEARS 2023*. Institute of Electrical and Electronics Engineers Inc., pp.885–889.

Banik, S., Sharma, N., Mangla, M., Mohanty, S.N. and Shitharth, S., (2022) LSTM based decision support system for swing trading in stock market. *Knowledge-Based Systems*, 239.

Barboza, F. and Altman, E., (2024) Predicting financial distress in Latin American companies: A comparative analysis of logistic regression and random forest models. *North American Journal of Economics and Finance*, 72.

Bitetto, A., Cerchiello, P. and Mertzanis, C., (2023) Measuring financial soundness around the world: A machine learning approach. *International Review of Financial Analysis*, 85.

Brenes, R.F., Johannssen, A. and Chukhrova, N., (2022) An intelligent bankruptcy prediction model using a multilayer perceptron. *Intelligent Systems with Applications*, 16.

Carmona, P., Climent, F. and Momparler, A., (2019) Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. *International Review of Economics and Finance*, 61, pp.304–323.

Che, W., Wang, Z., Jiang, C. and Abedin, M.Z., (2024) Predicting financial distress using multimodal data: An attentive and regularized deep learning method. *Information Processing and Management*, 614.

Chen, T.H., Lee, C.C. and Shen, C.H., (2022) Liquidity indicators, early warning signals in banks, and financial crises. *North American Journal of Economics and Finance*, 62.

Chen, Z., Chen, W. and Shi, Y., (2020) Ensemble learning with label proportions for bankruptcy prediction. *Expert Systems with Applications*, 146.

Cheng, C.H., Kao, Y.F. and Lin, H.P., (2021) A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes. *Applied Soft Computing*, 108.

Citterio, A. and King, T., (2023) The role of Environmental, Social, and Governance (ESG) in predicting bank financial distress. *Finance Research Letters*, 51, p.103411.

Çolak, M.S., (2021) A new multivariate approach for assessing corporate financial risk using balance sheets. *Borsa Istanbul Review*, 213, pp.239–255.

Ding, S., Cui, T., Bellotti, A.G., Abedin, M.Z. and Lucey, B., (2023) The role of feature importance in predicting corporate financial distress in pre and post COVID periods: Evidence from China. *International Review of Financial Analysis*, 90.

Dinh, D. V., Powell, R.J. and Vo, D.H., (2021) Forecasting corporate financial distress in the Southeast Asian countries: A market-based approach. *Journal of Asian Economics*, 74.

Du, X., Li, W., Ruan, S. and Li, L., (2020) CUS-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection. *Applied Soft Computing Journal*, 97.

Duan, J., (2019) Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction. *Journal of the Franklin Institute*, 3568, pp.4716–4731.

Elhoseny, M., Metawa, N. and El-hasnony, I.M., (2022) A new metaheuristic optimization model for financial crisis prediction: Towards sustainable development. *Sustainable Computing: Informatics and Systems*, 35.

Fan, M., Chen, X., Liu, B., Zhou, F., Gong, B. and Tao, R., (2023) An analysis of financial risk assessment of globally listed football clubs. *Heliyon*, 912.

Figlioli, B. and Lima, F.G., (2022) A proposed corporate distress and recovery prediction score based on financial and economic components. *Expert Systems with Applications*, 197.

Gabrielli, G., Melioli, A. and Bertini, F., (2023) High-dimensional Data from Financial Statements for a Bankruptcy Prediction Model. In: *Proceedings - 2023 IEEE 39th International Conference on Data Engineering Workshops, ICDEW 2023*. Institute of Electrical and Electronics Engineers Inc., pp.1–7.

Garcia, J., (2022) Bankruptcy prediction using synthetic sampling. *Machine Learning with Applications*, 9, p.100343.

Hajek, P. and Munk, M., (2024) Corporate financial distress prediction using the risk-related information content of annual reports. *Information Processing and Management*, 615.

Hassan, A. and Yousaf, N., (2022) Bankruptcy Prediction using Diverse Machine Learning Algorithms. In: *Proceedings - 2022 International Conference on Frontiers of Information Technology, FIT 2022*. Institute of Electrical and Electronics Engineers Inc., pp.106–111.

Hassanniakalager, A., Sermpinis, G. and Stasinakis, C., (2021) Trading the foreign exchange market with technical analysis and Bayesian Statistics. *Journal of Empirical Finance*, 63, pp.230–251.

Huang, B., Wei, J., Tang, Y. and Liu, C., (2021) Enterprise Risk Assessment Based on Machine Learning. *Computational Intelligence and Neuroscience*, 2021.

Huang, Y., Wang, Z. and Jiang, C., (2024) Diagnosis with incomplete multi-view data: A variational deep financial distress prediction method. *Technological Forecasting and Social Change*, 201.

Huang, Y.P. and Yen, M.F., (2019) A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing Journal*, 83.

Jabeur, S. Ben, Gharib, C., Mefteh-Wali, S. and Arfi, W. Ben, (2021) CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, 166.

Ben Jabeur, S. and Serret, V., (2023) Bankruptcy prediction using fuzzy convolutional neural networks. *Research in International Business and Finance*, 64.

Ben Jabeur, S., Stef, N. and Carmona, P., (2023) Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering. *Computational Economics*, 612, pp.715–741.

du Jardin, P., (2021a) Dynamic self-organizing feature map-based models applied to bankruptcy prediction. *Decision Support Systems*, 147.

du Jardin, P., (2021b) Forecasting bankruptcy using biclustering and neural network-based ensembles. *Annals of Operations Research*, 2991–2, pp.531–566.

Jiang, C., Zhou, Y. and Chen, B., (2023) Mining semantic features in patent text for financial distress prediction. *Technological Forecasting and Social Change*, 190.

Kadkhoda, S.T. and Amiri, B., (2024) A Hybrid Network Analysis and Machine Learning Model for Enhanced Financial Distress Prediction. *IEEE Access*, 12, pp.52759–52777.

Kaggle - Rubens Marques Chaves, (2023) *Financial Distress Prediction in Data Stream*. [online] Kaggle. Available at: https://www.kaggle.com/datasets/rubensmchaves/ml-fdp-ds [Accessed 27 Aug. 2024].

Kim, H.M., Bock, G.W. and Lee, G., (2021) Predicting Ethereum prices with machine learning based on Blockchain information. *Expert Systems with Applications*, 184.

Kim, S.Y. and Upneja, A., (2021) Majority voting ensemble with a decision trees for business failure prediction during economic downturns. *Journal of Innovation and Knowledge*, 62, pp.112–123.

Kottala, S.Y. and Sahu, A.K., (2024) Evaluating ergonomics and financial distress in manufacturing organization behavior: resiliency framework from operations and strategic management. *Learning Organization*, 315, pp.765–788.

Lall, R. and Robinson, T., (2022) The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning. *Political Analysis*, 302, pp.179–196.

Leng, A. and Sun, Y., (2024) The impact mechanism and breakthrough path of COVID-19 on enterprise financial distress: Evidence from China. *Economic Analysis and Policy*, 82, pp.16–31.

Liang, D., Tsai, C.F., Lu, H.Y. (Richard) and Chang, L.S., (2020) Combining corporate governance indicators with stacking ensembles for financial distress prediction. *Journal of Business Research*, 120, pp.137–146.

Liu, K., Zhou, J. and Dong, D., (2021) Improving stock price prediction using the long short-term memory model combined with online social networks. *Journal of Behavioral and Experimental Finance*, 30.

Liu, W., Fan, H., Xia, M. and Pang, C., (2022) Predicting and interpreting financial distress using a weighted boosted tree-based tree. *Engineering Applications of Artificial Intelligence*, 116.

Liu, X., Zhang, Y., Tian, M. and Chao, Y., (2023) Financial distress and jump tail risk: Evidence from China's listed companies. *International Review of Economics and Finance*, 85, pp.316–336.

Lohmann, C. and Möllenhoff, S., (2023) The bankruptcy risk matrix as a tool for interpreting the outcome of bankruptcy prediction models. *Finance Research Letters*, 55.

Lokanan, M. and Sharma, S., (2024) The use of machine learning algorithms to predict financial statement fraud. *British Accounting Review*.

Ma, Y., Han, R. and Wang, W., (2021) Portfolio optimization with return prediction using deep learning and machine learning. *Expert Systems with Applications*, 165.

Mai, F., Tian, S., Lee, C. and Ma, L., (2019) Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 2742, pp.743–758.

Matin, R., Hansen, C., Hansen, C. and Mølgaard, P., (2019) Predicting distresses using deep learning of text segments in annual reports. *Expert Systems with Applications*, 132, pp.199–208.

Murugan, M.S. and T, S.K., (2023) Large-scale data-driven financial risk management & analysis using machine learning strategies. *Measurement: Sensors*, 27.

Nyitrai, T. and Virág, M., (2019) The effects of handling outliers on the performance of bankruptcy prediction models. *Socio-Economic Planning Sciences*, 67, pp.34–42.

Papík, M. and Papíková, L., (2024) Automated Machine Learning in Bankruptcy Prediction of Manufacturing Companies. In: *Procedia Computer Science*. Elsevier B.V., pp.1428–1436.

Peng, Y.L. and Lee, W.P., (2021) Data selection to avoid overfitting for foreign exchange intraday trading with machine learning. *Applied Soft Computing*, 108.

Petropoulos, A. and Siakoulis, V., (2021) Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique. *Central Bank Review*, 214, pp.141–153.

Pinelis, M. and Ruppert, D., (2022) Machine learning portfolio allocation. *Journal of Finance and Data Science*, 8, pp.35–54.

Pokhrel, N.R., Dahal, K.R., Rimal, R., Bhandari, H.N., Khatri, R.K.C., Rimal, B. and Hahn, W.E., (2022) Predicting NEPSE index price using deep learning models. *Machine Learning with Applications*, 9, p.100385.

Putri, H.R. and Dhini, A., (2019) Prediction of Financial Distress: Analyzing the Industry Performance in Stock Exchange Market using Data Mining. In: *2019 16th International Conference on Service Systems and Service Management (ICSSSM)*. [online] IEEE, pp.1–5. Available at: https://ieeexplore.ieee.org/document/8887824/.

Qian, H., Wang, B., Yuan, M., Gao, S. and Song, Y., (2022) Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree. *Expert Systems with Applications*, 190.

Rahayu, D.S. and Suhartanto, H., (2020) Financial Distress Prediction in Indonesia Stock Exchange's Listed Company Using Case Based Reasoning Concept. In: *2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA)*. [online] IEEE, pp.1009–1013. Available at: https://ieeexplore.ieee.org/document/9101948/.

Rahayu, D.S. and Suhartanto, H., (n.d.) *Ensemble Learning in Predicting Financial Distress of Indonesian Public Company*.

Rahman, M.J. and Zhu, H., (2024) Predicting financial distress using machine learning approaches: Evidence China. *Journal of Contemporary Accounting and Economics*, 201.

Ramzan, S., (2023) Comparison of Financial Distress Prediction Models Using Financial Variables. In: *International Conference on Electrical, Computer and Energy Technologies, ICECET 2023*. Institute of Electrical and Electronics Engineers Inc.

Rathore, R.K., Mishra, D., Mehra, P.S., Pal, O., HASHIM, A.S., Shapi'i, A., Ciano, T. and Shutaywi, M., (2022) Real-world model for bitcoin price prediction. *Information Processing and Management*, 594.

Rubesam, A., (2022) Machine learning portfolios with equal risk contributions: Evidence from the Brazilian market. *Emerging Markets Review*, 51.

Samad, M.D., Abrar, S. and Diawara, N., (2022) Missing value estimation using clustering and deep learning within multiple imputation framework. *Knowledge-Based Systems*, 249.

Serrano-Cinca, C., Gutiérrez-Nieto, B. and Bernate-Valbuena, M., (2019) The use of accounting anomalies indicators to predict business failure. *European Management Journal*, 373, pp.353–375.

Shen, F., Liu, Y., Wang, R. and Zhou, W., (2020) A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment ☆. [online] 192, p.105365. Available at: https://doi.org/10.1016/j.knosys.2019.105365.

Son, H., Hyun, C., Phan, D. and Hwang, H.J., (2019) Data analytic approach for bankruptcy prediction. *Expert Systems with Applications*, 138.

Syed, A.M., Jreisat, A., Al-Mohamad, S., Khaki, A. and Ali, S.S., (2023) Financial Distress, Survival and Performance of Saudi Arabian Companies. In: *2023 International Conference on Sustainable Islamic Business and Finance, SIBF 2023*. Institute of Electrical and Electronics Engineers Inc., pp.37–41.

Tang, P., Tang, T. and Lu, C., (2024) Predicting systemic financial risk with interpretable machine learning. *North American Journal of Economics and Finance*, 71.

Titikkristanti, F. and Mahardika, I.P.A.B., (2023) Artificial Neural Network for Financial Distress Prediction on Energy Companies Listed in Indonesia. In: *2023 International Conference on Digital Business and Technology Management, ICONDBTM 2023*. Institute of Electrical and Electronics Engineers Inc.

Tölö, E., (2020) Predicting systemic financial crises with recurrent neural networks. *Journal of Financial Stability*, 49.

Tsai, C.F., Sue, K.L., Hu, Y.H. and Chiu, A., (2021) Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. *Journal of Business Research*, 130, pp.200–209.

Veganzones, D. and Séverin, E., (2018) An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, 112, pp.111–124.

Wang, D. ni, Li, L. and Zhao, D., (2022a) Corporate finance risk prediction based on LightGBM. *Information Sciences*, 602, pp.259–268.

Wang, G.J., Chen, Y., Zhu, Y. and Xie, C., (2024a) Systemic risk prediction using machine learning: Does network connectedness help prediction? *International Review of Financial Analysis*, 93.

Wang, J., Jiang, C., Zhou, L. and Wang, Z., (2024b) Assessing financial distress of SMEs through event propagation: An adaptive interpretable graph contrastive learning model. *Decision Support Systems*, 180.

Wang, J., Jiang, C., Zhou, L. and Wang, Z., (2024c) Representing and discovering heterogeneous interactions for financial risk assessment of SMEs. *Expert Systems with Applications*, 247.

Wang, S. and Chi, G., (2024) Cost-sensitive stacking ensemble learning for company financial distress prediction. *Expert Systems with Applications*, 255.

Wang, Y., Wang, C., Sensoy, A., Yao, S. and Cheng, F., (2022b) Can investors' informed trading predict cryptocurrency returns? Evidence from machine learning. *Research in International Business and Finance*, 62.

Wu, D., Ma, X. and Olson, D.L., (2022) Financial distress prediction using integrated Z-score and multilayer perceptron neural networks. *Decision Support Systems*, 159.

Xiao, J., Wen, Z., Jiang, X., Yu, L. and Wang, S., (2024a) Three-stage research framework to assess and predict the financial risk of SMEs based on hybrid method. *Decision Support Systems*, 177.

Xiao, J., Zhong, Y., Jia, Y., Wang, Y., Li, R., Jiang, X. and Wang, S., (2024b) A novel deep ensemble model for imbalanced credit scoring in internet finance. *International Journal of Forecasting*, 401, pp.348–372.

Yang, W., (2023) A Novel Intelligent Prediction Model for Corporate Distress Using Machine Learning and Support Vector Machine. In: *Proceedings - 2023 International Conference on Networking, Informatics and Computing, ICNETIC 2023*. Institute of Electrical and Electronics Engineers Inc., pp.162–166.

Yi, Z., Liang, Z., Xie, T. and Li, F., (2023) Financial risk prediction in supply chain finance based on buyer transaction behavior. *Decision Support Systems*, 170.

Yu, L. and Li, M., (2023) A case-based reasoning driven ensemble learning paradigm for financial distress prediction with missing data. *Applied Soft Computing*, 137.

Yu, L., Li, M. and Liu, X., (2024) A two-stage case-based reasoning driven classification paradigm for financial distress prediction with missing and imbalanced data. *Expert Systems with Applications*, 249.

Zeng, S. and Yang, W., (2020) Selection of Variables and Indicators in Financial Distress Prediction Model-Svm Method Based on Sparse Principal Component Analysis. In: *International Conference on Wavelet Analysis and Pattern Recognition*. IEEE Computer Society, pp.26–30.

Zhang, Z., Wu, C., Qu, S. and Chen, X., (2022) An explainable artificial intelligence approach for financial distress prediction. *Information Processing and Management*, 594.

Zhao, S., Xu, K., Wang, Z., Liang, C., Lu, W. and Chen, B., (2022) Financial distress prediction by combining sentiment tone features. *Economic Modelling*, 106.

Zhou, Y., Uddin, M.S., Habib, T., Chi, G. and Yuan, K., (2021) Feature selection in credit risk modeling: an international evidence. *Economic Research-Ekonomska Istrazivanja* , 341, pp.3064–3091.

Zoričák, M., Gnip, P., Drotár, P. and Gazda, V., (2020) Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets. *Economic Modelling*, 84, pp.165–176.

APPENDIX A: RESEARCH PROPOSAL

Exploring the Potential of Advanced Ensemble Tree Methods for classifying companies based on financial health: A Comparative Study using Financial Indicators

MITHUL MURUGAADEV

Research Proposal

May 2024

ABSTRACT

The ability to accurately predict financial distress is crucial for various stakeholders, including investors, creditors, and regulatory bodies. Traditional statistical methods for predicting financial distress, such as discriminant analysis and logistic regression, have limitations in handling complex, non-linear relationships, high-dimensional data, and scalability among other challenges. This thesis explores the application of advanced machine learning (ML) techniques for financial distress prediction, aiming to improve predictive accuracy and provide timely warnings of potential financial crises.

The research begins with a comprehensive review of existing literature on financial distress prediction models, highlighting the strengths and weaknesses of traditional approaches.

Several state-of-the-art ML algorithms, including random forests, regularized greedy forests, support vector machines, artificial neural networks, and extreme gradient boosting, are evaluated on a large dataset of accounting data, and financial ratios. The models are rigorously trained, validated, and tested using appropriate cross-validation techniques to ensure generalization and avoid overfitting.

The performance of the ML models is comprehensively assessed using various evaluation metrics, such as accuracy, F1 score, confusion matrix, and area under the receiver operating characteristic curve (AUC-ROC). In this in-depth analysis, the emphasis lies on identifying the optimal algorithm and methodologies for predicting financial distress.

Furthermore, the thesis investigates the interpretability and explainability of the ML models, an essential aspect of practical applications in the financial domain. The research findings contribute to the development of robust and accurate financial distress prediction models, enabling timely intervention and risk mitigation strategies. The thesis provides valuable insights for practitioners, regulators, and academic researchers in the fields of finance, risk management, and machine learning.

1. BACKGROUND

The exploration of financial distress has captivated researchers for an extended period. This phenomenon captures the state where a business faces substantial difficulties meeting its financial obligations. Several signs, such as diminishing revenue streams, falling profitability, and liquidity concerns, are indicative of it (Ramzan, 2023). Over time, there have been numerous advancements in the field of predicting financial distress, with bankruptcy prediction and credit scoring emerging as two widely utilized methods. In the past, determining credit risk mostly depended on the knowledge and experience of experts, who would render decisions using both their observations and set guidelines. Credit scoring models were created as a result of this traditional method, to systematize and improve the procedure by utilizing statistical methods and data analysis (Yang, 2023).

Since then, there have been advancements in statistical and accounting-based models for forecasting financial distress. Notable models include those pioneered by Altman, Ohlsen, and the Beneish M score, which gained popularity. Later on, statistical techniques such as logit, probit, and linear probability gained increased popularity in foreseeing corporate collapse. These models remain relevant today as researchers employ them in different methodologies to detect financial distress. However, they are not without limitations, such as reliance on assumptions and rule-based approaches, challenges with multicollinearity, outliers, missing data, and difficulties in handling large and diverse datasets(Gabrielli et al., 2023). This has opened the door for utilizing machine learning algorithms, which are adept at effectively addressing these limitations. As a result, there has been remarkable growth in this field

Numerous models have been developed to enhance prediction capabilities, ranging from conventional classifier algorithms to the creation of robust models employing artificial intelligence. A recent study used a hybrid model of combining PCA with ANN and demonstrated its effectiveness and high accuracy on a Polish companies' bankruptcy dataset (Adisa et al., 2019). A model built similarly to this one filters data characteristics and eliminates unnecessary information using a hybrid model with Sparse PCA and SVM (Zeng and Yang, 2020). An

alternative approach to bankruptcy prediction research suggests that company reports carry more utility than solely relying on ratios and other quantitative data. Specifically, findings indicate that auditor reports hold greater informative power for bankruptcy prediction compared to management statements, particularly when employing Convolutional Recurrent Neural Networks (CRNNs). This method has highlighted important data from the reports that are not usually captured by the numerical data (Matin et al., 2019).

A recent study showcased the efficacy of Case-Based Reasoning (CBR) by constructing a more resilient prediction model. This model utilized CBR and was developed using data from the Indonesian Stock Exchange spanning the years 2010 to 2016(Rahayu and Suhartanto, 2020b). Another study has shown to develop a financial crisis early warning system for manufacturing companies using a Random forest algorithm, enabling monitoring, risk identification, alerting and taking proactive control actions through the integrated machine learning model(Lyu and Xu, 2023).

A plethora of advanced and robust models are created that use quantitative and qualitative data. anticipating financial difficulties is critical to the health of the economy as well as to the financial security of individual people. Better decision-making is made possible by this predictive capacity for various stakeholders, including investors protecting their capital, entrepreneurs developing business plans, legislators forming economic policies, and governmental organizations organizing their activities. In brief, it is an indispensable instrument for all those concerned, enabling better-informed and efficient planning and decision-making procedures.

2. RELATED WORK

Financial distress prediction research spans various methodologies, from traditional ratios to cutting-edge machine learning. This collection showcases diverse approaches applied globally to enhance corporate finance distress prediction.

(Syed et al., 2023) delve into examining financial distress in Saudi Arabian companies by analyzing data from 178 firms. Through correlation and factor analysis, the researchers identified three crucial factors: leverage, liquidity, and efficiency, comprising eight critical ratios. Their findings emphasize the importance of these ratios in determining financial distress and influencing company survival and performance. Meanwhile, (Putri and Dhini, 2019)) has developed data mining models with logistic regression, C4.5 decision tree and ensemble classifiers among other algorithms, to predict financial distress of listed companies in the Indonesian stock exchange. The decision tree with boosting model performed best, with 94.61% accuracy, outperforming conventional statistical techniques. Another study spanning Indonesian energy sector companies, conducted by (Titikkristanti and Mahardika, 2023) proposed an ANN model to predict financial distress and develop an early warning system among energy companies listed on the Stock Exchange from 2018 to 2021. Motivated by the adverse effects of the COVID-19 pandemic on these companies' performance, they utilized four key financial ratios as input parameters for the model. The ANN model optimized to 4-5-1 configuration, has demonstrated enormous accuracy of 97.3% when tested on a sample of 58 companies. Similar studies have set a groundwork for much more exploration and refinement of advanced algorithms.

Building on the concept, (Zeng and Yang, 2020) proposed a hybrid financial distress prediction method combining sparse principal component analysis (SPCA) and support vector machines (SVM). Used SPCA to reduce the dimensionality of high-dimensional financial data by filtering redundant information in each group. The SPCA-SVM approach outperforms using all original indicators has performed well while reducing the misclassification of normal companies. A similar model called SRA-SVM was proposed in a recent study. The model combines Support Vector Machine (SVM) with stepwise regression analysis (SRA) to optimize feature selection and reduce multicollinearity issues and has exhibited good performance by leveraging time series forecasting (Yang, 2023). Addressing preprocessing of data, another study has shown the effectivity of DBN-

SVM model in predicting financial distress using accounting ratios of Taiwanese public firms. This hybrid model using deep belief network for feature extraction before classification by SVM, outperformed using SVM or DBN alone, highlighting the advantage of combining supervised and unsupervised techniques (Huang and Yen, 2019).

Expanding on the importance of preprocessing and handling imbalance of data, a research investigated bankruptcy prediction on Polish companies dataset. They addressed missing data and class imbalance issues through various imputation techniques and oversampling. Surprisingly, the simple mean imputation technique outperformed more complex imputation methods. The balanced bagging classifier with mean imputation achieved the highest accuracy of up to 98.23%, outperforming other models (Hassan and Yousaf, 2022).

Several studies have been conducted on the use of qualitative data in prediction. Notably, (Wei and Chen, 2022) have taken a different approach to predicting financial distress by integrating textual features from annual reports with financial indicator data. They utilized word2vec and ensemble learning methods to extract textual features and combined them with numerical financial indicators. Their E-CNN fusion model, employing convolutional neural networks, demonstrated improved prediction performance compared to using financial indicators alone. Various studies have showcased the potential of incorporating textual information from annual reports to enhance prediction capabilities (Matin et al., 2019).

Adding to the discourse, many studies have exhibited the use of Genetic algorithms for predicting financial distress. Authors have delved into building a GA-optimized neural network technique and the model has shown impressive accuracy of 84.7% with the test set (Yang and Yang, 2020). Empirical results show that the GA-BP model achieves higher prediction accuracy than the traditional BP models. As an alternative approach, Authors proposed a novel hybrid network analysis and machine learning model. By leveraging company networks based on financial indicator similarity and correlation, along with network-centric features, they achieved superior predictive capabilities by integrating them with original ratios. Their methodology leverages

network science to capture inter-firm relationships and dynamics for robust distress prediction (Kadkhoda and Amiri, 2024).

Expanding on the topic, (Rahayu and Suhartanto, 2020b) have taken another approach to predicting financial distress through Case-based reasoning. Their CBR approach retrieved similar past cases on financial ratios, reused class labels, and incorporated revise-retain steps to enhance the case base continuously. While having slightly lower accuracy than some machine learning models, CBR showed better performance in detecting the crucial minority distressed class, highlighting its potential for intelligent financial distress prediction by leveraging available corporate data repositories. Later (Rahayu and Suhartanto, 2020a) have experimented with predicting distress using ensembles – random forest and AdaBoost. This study showed that AdaBoost has better performance than random forests.

Further, there are several studies conducted depicting the strength of hybrid models and innovative cross-domain models in predicting financial distress. (Sun et al., 2021) presents innovative four-state multi-class financial distress prediction (FDP) models, advancing the theoretical system of multi-class FDP modeling. Notably, OVO-SVM demonstrates competitive performance compared with traditional statistical methods like multivariate discriminant analysis (MDA) and multinomial logit (MNLogit), particularly in predicting financial distress. Adding depth to the narrative, (Balachander et al., 2023) presents a novel approach, FCPFS-QDNN, for automated financial crisis prediction, combining feature subset selection and quantum deep neural networks. By integrating interactive search algorithms for feature selection, it achieves superior predictive accuracy compared to recent methods. There have been multiple advances in predicting financial hardship continuing to address challenges and building robust systems that are accurate and comprehensive.

This study integrates advanced methods across imputation, class imbalance handling, dimensionality reduction, and utilizes advanced classifiers such as regularized greedy forest and

ensemble methods, alongside comprehensive evaluation techniques, to enhance financial distress prediction.

3. RESEARCH QUESTIONS

This research tries to answer the following questions:

1. What machine learning techniques are most effective in predicting financial distress in companies using high-dimensional accounting and financial data?
2. What are the comparative performances of various ensemble tree methods and other advanced machine learning algorithms in financial distress prediction?
3. What evaluation metrics provide the most comprehensive assessment of machine learning models used for financial distress prediction?

4. AIM & OBJECTIVES

The research is conducted with the goal of developing a robust and comprehensive machine learning framework for predicting financial distress in companies using high-dimensional accounting and financial data, incorporating various techniques to address class imbalance, dimensionality reduction, and ensemble modeling.

The research objectives are formulated based on the aim of this study, which are as follows:

- To preprocess high-dimensional data using suitable techniques for noisy data, missing values, and outliers.
- To Investigate and implement appropriate class imbalance techniques.
- To build predictive models to identify the most accurate and high-performing model to classify distressed companies based on financial and accounting data.
- To evaluate the performances of the classifiers using comprehensive evaluation metrics and techniques.

5. SIGNIFICANCE OF STUDY

This study holds significant implications for academia, industry, and regulatory bodies alike. By employing machine learning techniques and frameworks to predict financial distress, it aims to enhance financial risk management practices and ensure the sustainability of businesses.

Early detection and proactive mitigation of financial distress are crucial for safeguarding investments and maintaining business continuity. Through the development of robust predictive models, this study empowers decision-makers to make informed strategic decisions, optimize resource allocation, and navigate financial challenges effectively.

The interdisciplinary nature of this research fosters collaboration and innovation across finance, data science, and regulatory domains. By facilitating knowledge exchange and continuous improvement, it drives progress and adaptation in an evolving global landscape.

Lastly, this study expands theoretical understanding and empirical evidence in financial risk management

6. SCOPE OF STUDY

The scope of this thesis work is defined as follows:

- The thesis will focus on the development and evaluation of machine learning models and methodologies for financial distress prediction using high-dimensional accounting and financial data.
- The data utilized in this research are derived from accounting and financial statements, including balance sheets, income statements, and cash flow statements.

7. RESEARCH METHODOLOGY

The use of machine learning models in addressing financial distress prediction has become prominent as it offers sophisticated tools to analyze vast amounts of financial data, enhancing the accuracy and reliability of predictions. There are various studies conducted exploring the usage and combination of methods spanning across feature selection, preprocessing, and model building along with ensembles (Tsai et al., 2021; Novaldo et al., 2023). A recent study developed a hybrid SRA-SVM model for distress prediction(Yang, 2023). Another research has utilized SPCA (sparse PCA) for variable selection and dimensionality reduction, further building a hybrid model with SPCA-SVM, the authors have proven that the model has achieved high accuracy and performance standards (Zeng and Yang, 2020). Although, most of the relevant studies state multicollinearity, overfitting, and noisy data are constant challenges faced in predicting financial distress.

7.1 Dataset Description:

This dataset contains real-world financial and accounting data of Brazilian firms sourced from the Brazilian Securities and Exchange Commission (CVM), which are structured quarterly in a non-stationary way. The dataset spans ten years from 2011 to 2020 and consists of 23,834 records from 905 distinct corporations, each characterized by 84 indicators. Just 651 businesses encountered financial trouble, whilst the majority of businesses exhibited no financial difficulties. The data shows a significant imbalance, with 2.73% of the data pertaining to enterprises experiencing financial hardship, and 97.27% not. This dataset is collected from Kaggle, shared by Rubens Marques Chave (Rubens Marques Chave, 2024)

This dataset comprises accounting and finance data from the balance sheet, income statement, and cash flow statement, along with derived ratios and metrics. All the variables can be categorized under these respective headers.

- Assets columns: [A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12]
- Liabilities columns: [A13, A14, A15, A16]
- Equity: [A17, A18, A19]
- Loans and debts: [A20, A21]
- Income and earnings: [A22, A23, A24, A25, A26, A27, A28, A29, A30, A31]

- Cash flow: [A32, A33, A34]
- Shares: [A35]

Ratios

- solvency ratios: [A36, A37, A38, A39, A40, A41, A42, A43, A44]
- efficiency ratios: [A45, A46, A47, A48, A49, A50, A51, A52, A53, A54, A55, A56]
- profitability ratios: [A57, A58, A59, A60, A61, A62, A63, A64, A65, A66]
- leverage ratios: [A71, A72, A73]
- Growth rates: [A74, A75, A76, A77, A78, A79, A80, A81]
- Per share metrics: [A82, A83, A84]
- Label: 0: normal situation; 1: financial distress situation

However, this study focuses on using ratios as the input, as these comprehensive ratios are derived from data available in the balance sheet, income statement, and cash flow statement.

7.2 Dataset Preprocessing:

Effective data preprocessing is crucial for accurate financial distress prediction using machine learning models. Addressing the importance of the preprocessing of financial data, studies exploring imputation techniques and handling class imbalance techniques have comprehensively discussed its effect on the performance and accuracy of the models (Cheng et al., 2019).

Researchers have also proposed hybrid models along with conventional methods and models for addressing preprocessing challenges. Notably among these is ADASVM-TW, which showcases the potential of hybrid techniques in improving predictive accuracy (Sun et al., 2020). This study focuses on addressing the following challenges by exploring various methods:

- Imputation: Missing data can introduce bias and hinder the generalization capabilities of algorithms, necessitating effective imputation techniques. Hence, they are treated with imputation techniques. This study focuses on exploring the accuracy using KNN-based imputation, mean imputation. A detailed study comparing various methods stated mean imputation has resulted in higher accuracy of the models (Hassan and Yousaf, 2022).

- Imbalance data: Addressing the challenges posed by heavily imbalanced datasets, this study employs Cluster-based SMOTE. Notably, a recent study proposed KNSMOTE, which has demonstrated superior results over traditional methods, particularly in the context of high dimensional data(Xu et al., 2021).
- Grouped PCA: To facilitate dimensionality reduction, ratios are categorized under relevant headers and then processed with Principal Component Analysis (PCA). A similar methodology has been followed in the study developing the SPCA-SVM model, which stated that grouping significant variables and removing noise variables has resulted in higher performance (Zeng and Yang, 2020).
- Transformation: numerical columns are transformed using the standard scaler method.

7.3 Models:

This study involves comparing various models for the classification of high-dimensional financial data. The suitable models used are as follows:

- Regularized Greedy Forest (RGF) - This technique employs a fully corrective regularized greedy search to construct a decision forest. RGF addresses the issue of the lack of direct regularization in the gradient-boosting decision tree (GBDT) method. By incorporating regularization directly into the tree-building process, RGF aims to enhance model generalization and prevent overfitting. Recent research has shown the impressive accuracy of RGF compared with other models in the detection of DGA (Mitsuhashi et al., 2023).
- Random forest (RF): The Random Forest is like a team of decision-making trees that come together to form a diverse forest. By repeatedly training on different parts of the data, this forest of trees avoids being too similar. This teamwork results in a group of trees that are less affected by individual quirks in the data and more stable in their predictions (Ramzan, 2023). Several studies have highlighted the robustness of RF in predicting financial distress and other complex phenomena.
- Support Vector Machine (SVM): The Support Vector Machine is a supervised learning algorithm commonly used for classification and regression tasks. Despite often having lower performance compared to other advanced models, SVM can achieve high accuracy

when combined with dimensionality reduction techniques. It has been seen to perform high accuracy when utilized with dimensionality reduction techniques (Huang and Yen, 2019; Zeng and Yang, 2020; Yang, 2023).

- Extreme Gradient Boosting (XGBoost, XGB): XGBoost stands as a powerful tool in supervised learning, adept in handling both regression and classification tasks. It seamlessly blends gradient descent with tree ensemble learning techniques. Employing a methodical approach, it iteratively enhances model performance by incorporating new trees, ultimately refining prediction accuracy and fine-tuning specified objective functions. In a comprehensive study comparing the models, XGB has been stated to have the highest performance comparatively (Huang and Yen, 2019).
- Artificial Neural Network (ANN): An Artificial Neural Network mirrors the intricate connections found in the human brain, forming a computational framework. It's comprised of nodes, akin to neurons, structured into layers for input, hidden, and output, allowing for complex data processing akin to the brain's functions. Researchers have demonstrated the use of ANN for building robust models with great accuracy both as standalone models and in hybrid configurations, making them versatile tools in various predictive tasks ((Tsai et al., 2021; Titikkristanti and Mahardika, 2023).

7.4 Evaluation Metrics:

These are the following methods used for evaluating the models and their results.

- Accuracy: Accuracy is a straightforward metric that measures the proportion of correct predictions (both true positives and true negatives) out of all predictions made. It is widely used due to its simplicity and effectiveness in providing a quick overview of a model's performance, especially when the dataset is balanced.
- F1 score: The F1 score is particularly valuable for understanding the impact and consequences of false positives and false negatives. F1 is highly used in the evaluation of classification models along with precision and recall (Novaldo et al., 2023)

- AUC-ROC: The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a robust metric, especially suited for evaluating models on imbalanced datasets. It measures the model's ability to distinguish between classes, providing insights into the trade-off between true positive rates and false positive rates (Tsai et al., 2021)
- Confusion matrix: The confusion matrix offers a granular view of a model's accuracy, detailing correct predictions (true positives and true negatives) as well as errors (false positives and false negatives), enabling an understanding of the types and costs of misclassifications. Many studies have made use of this since it provides a detailed perspective that makes it easier to determine the expenses related to various kinds of errors, allowing for a more complex assessment of the model's advantages and disadvantages (Qian et al., 2022).

These methods help in identifying the overall correctness, diagnosing errors and class-specific performances of the models.

8. REQUIRED RESOURCES

Resources required are as furnished below:

8.1 Hardware requirements:

- A laptop/desktop computer with internet connectivity, capable of web browsing.
- RAM: 8GB
- Graphics Card: Integrated GPU (Intel Iris Xe Graphics or equivalent)
- Processor: 6th Generation Intel Core i5 or equivalent (4 cores, 8 threads)
- Storage: 256GB SSD

8.2 Software requirements:

- Programming Language:
    - Python (version 3.8 or later)
- Machine Learning Libraries:

    - TensorFlow (version 2.2 or later)
    - Scikit-learn (version 1.0.1 or later)
    - Statsmodels (version 0.12.0 or later)
- Visualization Tools:
    - Matplotlib (version 3.6.0 or later)
    - Seaborn (version 0.12.0 or later)
- Data Processing Libraries:
    - Pandas (version 2.0.0 or later)
    - NumPy (version 1.23.0 or later)
- Development Environment:
    - Jupyter Notebook (version 6.5.3 or later) for an interactive coding and data exploration environment.
- Version Control:
    - Git for (version 2.42.0 or later) tracking changes in code and data throughout the research.

Additional advanced software tools and libraries might be necessary depending on the complexity of the models being developed. This configuration ensures the essential resources, including data, software, hardware, and expertise, to explore and investigate the advantages and challenges of applying machine learning techniques for financial distress prediction.

9. RESEARCH PLAN

The project plan has been furnished below:

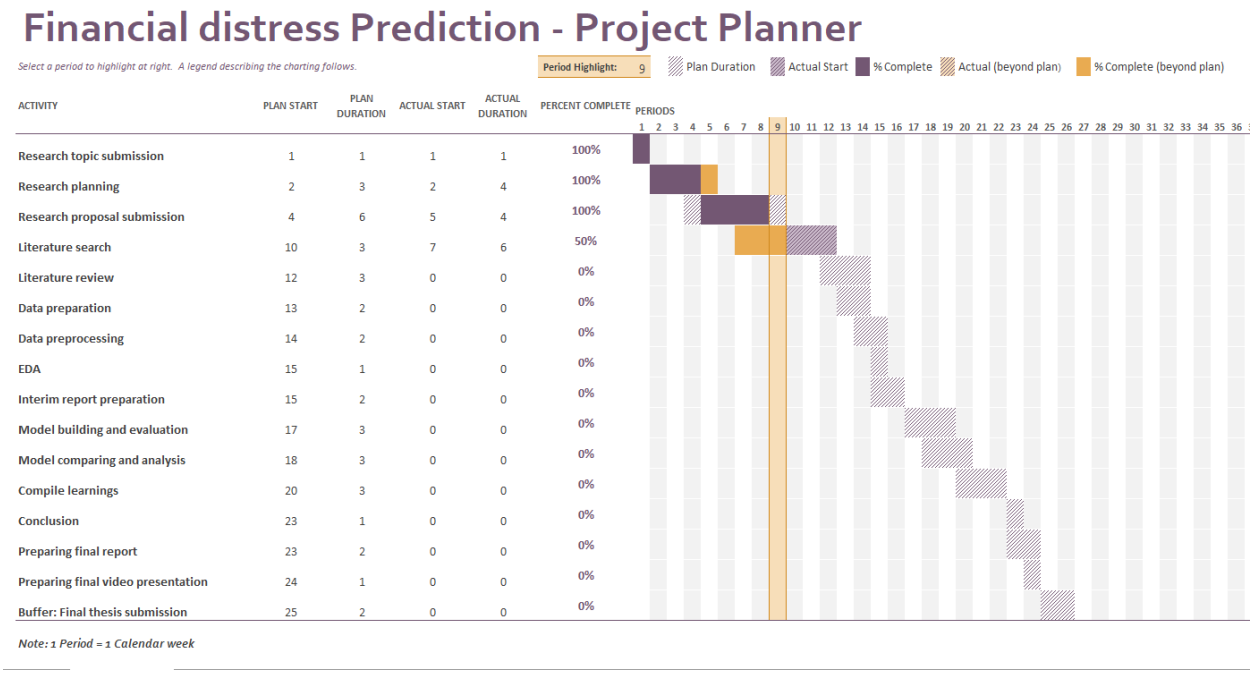

Figure 1: Research plan

REFERENCES

Adisa, J.A., Ojo, S.O., Owolawi, P.A. and Pretorius, A.B., (2019) Financial Distress Prediction: Principle Component Analysis and Artificial Neural Networks. In: 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC). [online] IEEE, pp.1–6. Available at: https://ieeexplore.ieee.org/document/9015884/.

Balachander, T., Akhlaq, N., Bansal, R., Vasani, S.A., Singh, K. and Mannar, B.R., (2023) Financial Crisis Prediction using Feature Subset Selection with Quantum Deep Neural Network. In: Proceedings of the 2023 2nd International Conference on Electronics and Renewable Systems, ICEARS 2023. Institute of Electrical and Electronics Engineers Inc., pp.885–889.

Cheng, C.H., Chan, C.P. and Sheu, Y.J., (2019) A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. Engineering Applications of Artificial Intelligence, 81, pp.283–299.

Gabrielli, G., Melioli, A. and Bertini, F., (2023) High-dimensional Data from Financial Statements for a Bankruptcy Prediction Model. In: Proceedings - 2023 IEEE 39th International Conference on Data Engineering Workshops, ICDEW 2023. Institute of Electrical and Electronics Engineers Inc., pp.1–7.

Hassan, A. and Yousaf, N., (2022) Bankruptcy Prediction using Diverse Machine Learning Algorithms. In: Proceedings - 2022 International Conference on Frontiers of Information Technology, FIT 2022. Institute of Electrical and Electronics Engineers Inc., pp.106–111.

Huang, Y.P. and Yen, M.F., (2019) A new perspective of performance comparison among machine learning algorithms for financial distress prediction. Applied Soft Computing Journal, 83.

Kadkhoda, S.T. and Amiri, B., (2024) A hybrid Network Analysis and Machine Learning model for Enhanced Financial Distress Prediction. IEEE Access.

Lyu, J. and Xu, H., (2023) Design of Modern Enterprise Financial Crisis Warning System Based on Random Forest Algorithm. In: 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN). pp.679–684.

Matin, R., Hansen, C., Hansen, C. and Mølgaard, P., (2019) Predicting distresses using deep learning of text segments in annual reports. Expert Systems with Applications, 132, pp.199–208.

Mitsuhashi, R., Jin, Y., Iida, K., Shinagawa, T. and Takai, Y., (2023) Detection of DGA-based Malware Communications from DoH Traffic Using Machine Learning Analysis. In: Proceedings - IEEE Consumer Communications and Networking Conference, CCNC. Institute of Electrical and Electronics Engineers Inc., pp.224–229.

Novaldo, W., Gunawan, A.A.S. and Ibrahim, M.A., (2023) Design of Company Financial Health Prediction System with Extra Trees Method Based on Fundamental Analysis. In: Proceeding - COMNETSAT 2023: IEEE International Conference on Communication, Networks and Satellite. Institute of Electrical and Electronics Engineers Inc., pp.232–238.

Putri, H.R. and Dhini, A., (2019) Prediction of Financial Distress: Analyzing the Industry Performance in Stock Exchange Market using Data Mining. In: 2019 16th International Conference on Service Systems and Service Management (ICSSSM). pp.1–5.

Qian, H., Wang, B., Yuan, M., Gao, S. and Song, Y., (2022) Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree. Expert Systems with Applications, 190.

Rahayu, D.S. and Suhartanto, H., (2020a) Ensemble Learning in Predicting Financial Distress of Indonesian Public Company. In: 2020 8th International Conference on Information and Communication Technology (ICoICT). pp.1–5.

Rahayu, D.S. and Suhartanto, H., (2020b) Financial Distress Prediction in Indonesia Stock Exchange's Listed Company Using Case Based Reasoning Concept. In: 2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA). [online] IEEE, pp.1009–1013. Available at: https://ieeexplore.ieee.org/document/9101948/.

Ramzan, S., (2023) Comparison of Financial Distress Prediction Models Using Financial Variables. In: International Conference on Electrical, Computer and Energy Technologies, ICECET 2023. Institute of Electrical and Electronics Engineers Inc.

Rubens Marques Chave, (2024) Financial Distress Prediction in Data Stream | Kaggle. [online] Kaggle. Available at: https://www.kaggle.com/datasets/rubensmchaves/ml-fdp-ds [Accessed 19 May 2024].

Sun, J., Fujita, H., Zheng, Y. and Ai, W., (2021) Multi-class financial distress prediction based on support vector machines integrated with the decomposition and fusion methods. Information Sciences, 559, pp.153–170.

Sun, J., Li, H., Fujita, H., Fu, B. and Ai, W., (2020) Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Information Fusion, 54, pp.128–144.

Syed, A.M., Jreisat, A., Al-Mohamad, S., Khaki, A. and Ali, S.S., (2023) Financial Distress, Survival and Performance of Saudi Arabian Companies. In: 2023 International Conference on Sustainable Islamic Business and Finance, SIBF 2023. Institute of Electrical and Electronics Engineers Inc., pp.37–41.

Titikkristanti, F. and Mahardika, I.P.A.B., (2023) Artificial Neural Network for Financial Distress Prediction on Energy Companies Listed in Indonesia. In: 2023 International Conference on Digital Business and Technology Management, ICONDBTM 2023. Institute of Electrical and Electronics Engineers Inc.

Tsai, C.F., Sue, K.L., Hu, Y.H. and Chiu, A., (2021) Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. Journal of Business Research, 130, pp.200–209.

Wei, X. and Chen, Y., (2022) Early Warning Model for Financial Risks of Listed Companies Based on Machine Learning. In: Proceedings - 2022 International Conference on Machine Learning and Intelligent Systems Engineering, MLISE 2022. Institute of Electrical and Electronics Engineers Inc., pp.473–477.

Xu, Z., Shen, D., Nie, T., Kou, Y., Yin, N. and Han, X., (2021) A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data. Information Sciences, 572, pp.574–589.

Yang, W., (2023) A Novel Intelligent Prediction Model for Corporate Distress Using Machine Learning and Support Vector Machine. In: Proceedings - 2023 International Conference on Networking, Informatics and Computing, ICNETIC 2023. Institute of Electrical and Electronics Engineers Inc., pp.162–166.

Yang, Y. and Yang, C., (2020) Research on the application of ga improved neural network in the prediction of financial crisis. In: Proceedings - 2020 12th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2020. Institute of Electrical and Electronics Engineers Inc., pp.625–629.

Zeng, S. and Yang, W., (2020) Selection of Variables and Indicators in Financial Distress Prediction Model-Svm Method Based on Sparse Principal Component Analysis. In: International Conference on Wavelet Analysis and Pattern Recognition. IEEE Computer Society, pp.26–30.