PREDICTIVE FINANCIAL DISTRESS ANALYSIS: ENHANCING RISK MITIGATION AND DECISION-MAKING USING MACHINE LEARNING.

By

Mghenyi Mike Kililo

IN16/00036/20

And

Stephen Kinuthia Kinyuru

IN16/0004/20

DEPARTMENT OF

SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY

A Project Proposal Submitted to the School of Information Science and Technology

for the Study Leading to a Project in Partial Fulfilment of the

Requirements for the Award of the Degree of Bachelor of

of Kisii University.

September, 2023

DECLARATION

We have done this work and neither is it copied nor stolen from any previously done work.

Signature: ……. Date: ……

Mike Kililo Mghenyi

Signature: ……. Date: ……

Kinyuru Stephen Kinuthia

This proposal has been submitted for examination with my approval as University Supervisor

Signature……...Date: ……...

Teresa Kwamboka Abuya

Kisii University, Kenya

DEDICATION

This work is dedicated to all aspiring entrepreneurs and business persons who wish to further the development of mankind as they better themselves in their path to financial freedom, while ensuring all they have built is well safeguarded from activities and events that would result in bankruptcy and financial doom.

ACKNOWLEDGEMENT

We would like to acknowledge the university for providing internet resources and the library staff who have been helpful. We would also like to acknowledge some of our friends and supervisor, Mrs. Teresa Abuya, who provided great insight to this project. We would also like to appreciate the founders of Google Scholar, YouTube, PubMed and other sites for providing huge spaces for gaining valuable resources and knowledge.

TABLE OF CONTENTS

DECLARATION II
DEDICATION III
ACKNOWLEDGEMENT IV
TABLE OF CONTENTS V
LIST OF FIGURES VII
LIST OF TABLES VIII
ABBREVIATIONS AND ACCROYNMS IX
ABSTRACT X
CHAPTER ONE 1
1.0 INTRODUCTION 1
1.1 BACKGROUND 1
1.2 PROBLEM STATEMENT 4
1.3 OBJECTIVES OF THE STUDY 5
1.4 RESEARCH QUESTIONS 5
1.5 JUSTIFICATION 6
1.6 SCOPE 6
CHAPTER TWO 7
2.0 LITERATURE REVIEW. 7
2.1 INTRODUCTION 7
2.2 EXISTING MEASURES FOR FINANCIAL DISTRESS PREDICTION 7
2.2.1 STATISTICAL APPROACHES 7
2.2.2 MACHINE LEARNING APPROACHES 9
2.3 CHALLENGES FACED BY EXISTING APPROACHES 11
CHAPTER THREE 15
3.0 METHODOLOGY 15
3.1 INTRODUCTION 15
3.2 DATASET DESCRIPTION 16
3.3 THEORETICAL DESCRIPTION OF CLASSIFIERS 19
3.3.1 DECISION TREES 19
3.3.2 RANDOM FOREST 19
3.3.3 SUPPORT VECTOR MACHINES 20

3.4 DATA PRE-PROCESSING 21
3.5 EVALUATION METRICS 22
3.6 PSEUDOCODE 24
3.7 FEATURE SELECTION 25
3.8 EXPERIMENTAL SET UP AND DESIGN 27
REFERENCES 28

LIST OF FIGURES

Figure 1.1.1: U.S. companies that filed for bankruptcy due to financial distress……2

Figure 3.1.1: Flowchart Diagram…….15

Figure 3.3.2.1 How Random Forest and Decision Trees work……20

Figure 3.5.1 How Confusion Matrix Works…….23

Figure 3.7.1 PCA flowchart…….26

LIST OF TABLES

Table 2.3.1 Challenges Faced by existing approaches……11

Table 3.2.1 Dataset attributes……...16

ABBREVIATIONS AND ACCROYNMS

FD: Financial Distress

API: Application Programming Interface

SVM: Support Vector Machine

MDA: Multiple Discriminant Analysis

DA: Discriminant Analysis

SME: Small Management Enterprise

NN: Neural Network

DNN: Dense Neural Network

LR: Logistic Regression

XGB: Extreme Gradient Boosting

PCA: Principal Component Analysis

ABSTRACT

The ever-evolving landscape of financial markets demands innovative approaches to risk management and decision-making. This study delves into the realm of predictive financial distress analysis, leveraging the power of machine learning techniques. By harnessing advanced algorithms such as Support Vector Machines, Random Forests and Decision Trees coupled with extensive historical financial data, this research aims to enhance risk mitigation strategies and bolster decision-making processes within the financial sector. Through a comprehensive examination of key financial indicators and their predictive power, this research project aims to unveils insights that will enable early identification of companies on the brink of financial distress. The integration of machine learning models with a web application using APIs not only refines predictive accuracy but also provides actionable intelligence for stakeholders, empowering them to implement pre-emptive measures and navigate turbulent economic climates with greater confidence. This research stands as a pivotal contribution to the domain of financial risk management, offering a forward-looking approach to fortify the stability and resilience of financial enterprises in an era of dynamic economic uncertainties.

CHAPTER ONE

1.0 INTRODUCTION

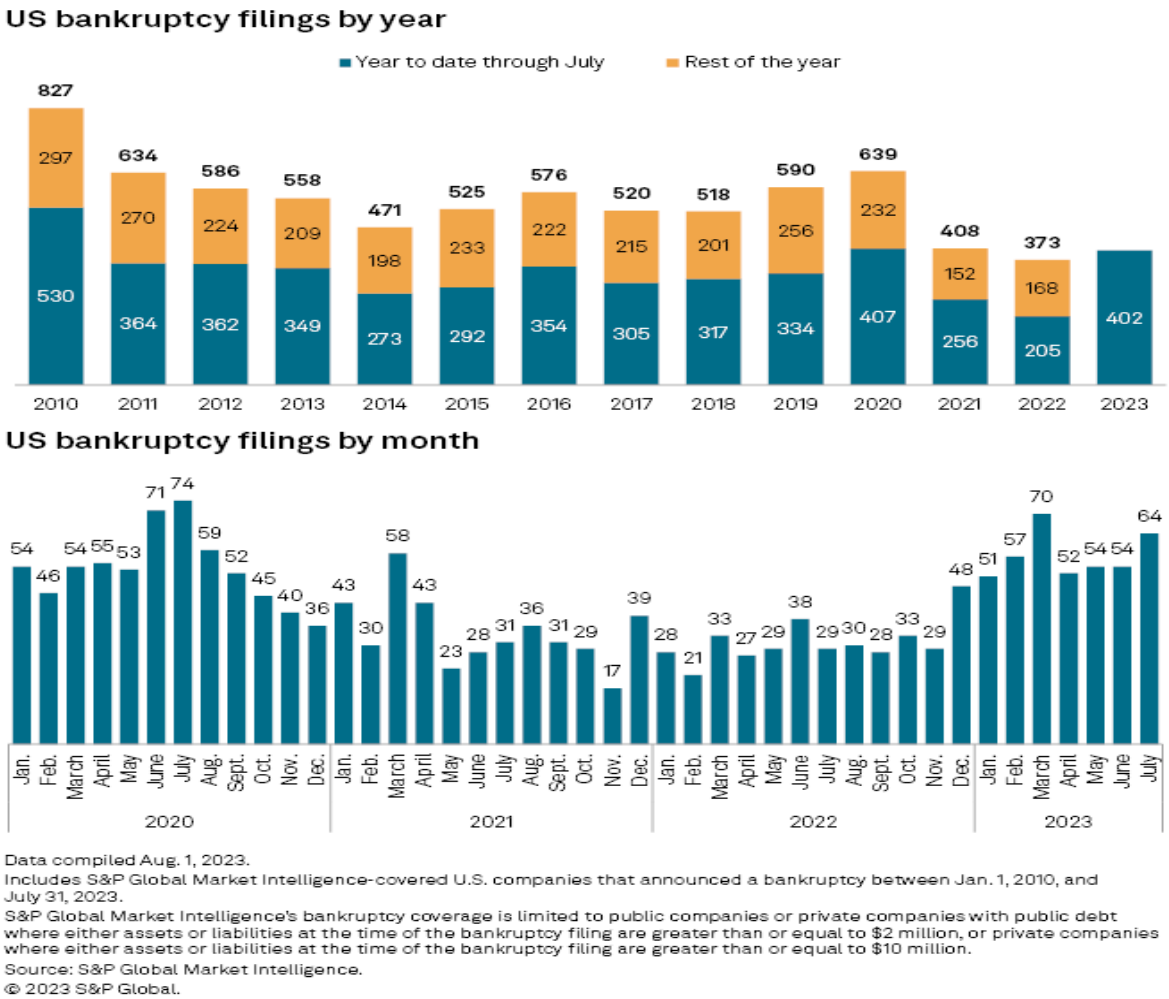
1.1 BACKGROUND

Financial distress is a situation in which a company or individual is unable to meet its set financial obligations such as paying bills, making loan payments, or meeting other financial commitments. Hamid et al., (2022) considers financial distress in an organisation as a highly risky condition that could lead to bankruptcy, with whom Altman argued that the public declaration of bankruptcy is the sole criterion of financial distress (Olsen et al., 2022), foreclosure and other negative consequences. Financial distress can be caused by a variety of factors, including: Economic downturns, Poor management, Fraud, Unexpected expenses, Loss of income and High debt levels.

Boomey., (2020) notes that financial distress can affect major organisations such as Brooks Brothers, a franchise firm in the clothing industry, that filed for financial bankruptcy which resulted in the closure of its nearly 200 stores worldwide in 2020. In June 2023, indoor agriculture tech company AeroFarms filed for Chapter 11 bankruptcy protection due to lowering interests from investors in recent years which can be attributed to capital headwinds such as rising interest rates and economic recessions (Casey 2023).

During the Covid-19 pandemic in Indonesia since March 2020, all sectors including the economic sector were not spared from financial losses (Rizal, 2020). The sectors that experienced the most pressure were the infrastructure sector and the trade, services, and investment sector, which showed a decline in performance between 2020 and 2021 (Putra, 2021; Ulya, 2020). Sidik (2020) confirms that companies during this time faced a decrease in corporate income and disrupted cash flows as they incurred operating costs while the purchasing power of their customers declined. This period increases the potential for companies to be faced by financial distress as they struggle to break even with their profits. In such a phase, preventive measures need to be undertaken to avoid financial distress.

The figures below from S&P Global Market Intelligence imply that in the U.S. alone, 64 corporate bankruptcies were filed in July, the largest monthly total since March 2023 and more filings than in any single month in 2021 and 2022. Filings in the first seven months of 2023 surpassed total filings for the previous year and the total fillings of the first seven months between 2013 till 2019.



Data compiled Aug. 1, 2023.
Includes S&P Global Market Intelligence-covered U.S. companies that announced a bankruptcy between Jan. 1, 2010, and July 31, 2023.
S&P Global Market Intelligence's bankruptcy coverage is limited to public companies or private companies with public debt where either assets or liabilities at the time of the bankruptcy filing are greater than or equal to $2 million, or private companies where either assets or liabilities at the time of the bankruptcy filing are greater than or equal to $10 million.
Source: S&P Global Market Intelligence.

Figure 1.1.1: U.S. companies that filed for bankruptcy due to financial distress.

Financial distress can affect anyone, regardless of income, age, or background. However, some groups of people are more likely to experience financial distress than others. These include:

- Low-income households: The American Psychological Association, (2018) suggests that low-income households often have limited financial resources and are more likely to be affected by unexpected expenses, such as loss of jobs or huge medical bills.
- People of colour: People of colour are more likely to live in poverty and experience financial distress than white people. This is due to a number of factors such as systemic racism and discrimination. (American Psychological Association, 2023).
- Single parents: According to the Center for American Progress, (2021), single parents often face unique financial challenges, such as child care costs and lack of a second income.
- People with disabilities: Disabled people are more likely to experience poverty and unemployment than people without disabilities. This can make it difficult to meet basic financial needs. (National Bureau of Economic Research, 2022)
- Older adults: Older adults may face financial challenges such as retirement savings that are inadequate to meet their needs, long-term care costs, and medical bills. (Economic Policy Institute, 2020)

In addition to these groups, anyone can experience financial distress due to unexpected events, such as a job loss, medical emergency, or natural disaster. Valenzuela et al.,(2022) notes that financial distress can also have a significant impact on people's lives and could lead to stress, anxiety, depression, and other health problems since it makes it difficult to afford housing, food, and other essential needs resulting in a decline in one’s overall life satisfaction.

1.2 PROBLEM STATEMENT

Financial distress is referred to as a situation where a firm is unable to generate sufficient funds to meet its financial obligations as at when they are due (Ikpesu & Eboiyehi, 2018). Financial distress is a global phenomenon that affects businesses and people worldwide and affected parties such as investors and organisations face financial losses and loss of skilled labour due to layoffs. Lexova and Khan (2023) from S&P Global Market Intelligence recorded 402 corporate bankruptcy filings in the U.S. alone by the end of July, which surpassed total filings for 2022 and was on par with the full tally of 2021's total filings as shown in Figure 1. This inadvertently results in loss of potential tax revenue to the government, joblessness which could lead to depression and loss of skilled labour and loss of potential financial gains by corporate organisations.

Hamori et al., (2018) used bagging, boosting and neural networks machine learning methods for credit risk management so as to help companies avoid financial distress while using a dataset with over 25000 observations. He found out that the bagging outperformed all other methods achieving an F-score of 71%, however in his study, despite an impressive performance, no dimensionality reduction techniques such as PCA were used within the huge dataset which would have resulted in higher performance. Olsen et al., (2022) used integrated Z-score and multilayer perceptron neural networks to predict FD and achieved an accuracy of 86.54% without the use of dimensionality reduction techniques.

Thus, in our study, we aim to predict financial distress using a model built with SVM, Decision Trees, Random Forest while using Principal Component Analysis. The model will then be integrated to a web application built with Fast API and ReactJS for the backend and frontend side respectively.

1.3 OBJECTIVES OF THE STUDY

1.3.1 MAIN OBJECTIVES

This study aims to help companies identify factors that would lead them to financial distress which could result in bankruptcy and huge financial losses using a web application integrated with machine learning techniques and APIs.

1.3.2 SPECIFIC OBJECTIVES

1. To analyse the factors that attribute to financial distress in companies.

2. To analyse previously used approaches in the prediction of FD.

3. To build a functional machine learning model and integrate it to a web application while utilising Fast API.

4. To test the performance of the web app.

1.4 RESEARCH QUESTIONS

1. What are the key factors contributing to financial distress in companies across various industries and economic contexts?
2. What are the strengths and limitations of previously employed approaches for predicting financial distress, and how can they be improved or extended in light of evolving data sources and analytical techniques?
3. How can a functional machine learning model be constructed and integrated into a web application using Fast API to facilitate real-time prediction and decision-making in the context of financial distress?
4. What is the performance and usability assessment of the web application in effectively predicting financial distress, and how does it compare to existing methods or tools for risk assessment and mitigation?

1.5 JUSTIFICATION

The main aim of this project is to help companies and industries reduce their chances of being affected by financial distress. This project aims to use robust machine learning techniques to assist companies in detecting early signs of financial distress which will help them in avoiding heavy financial losses and loss of skilled labour. As suggested by Lexova and Khan (2023), with over 400 companies filing for bankruptcy by the end of July 2023, this inadvertently results in loss of skilled workforce due to layoffs, loss of investors' trust and lower tax income to the government. As noted by Valenzuela et al.,(2022), FD can also have a significant impact on people's lives and could lead to stress, anxiety, depression, and other health problems since it makes it difficult to afford housing, food, and other essential needs resulting in a decline in one's overall life satisfaction. By being able to detect FD in its early stages, companies can be able to retain their skilled workforce, gain the trust of old and new investors and help in reduction of high joblessness rates.

1.6 SCOPE

Our study will focus on detecting financial distress while using data from companies that faced financial distress between 2014 and 2023. The companies' data to be used is not limited to any industry.

CHAPTER TWO

2.0 LITERATURE REVIEW.

2.1 INTRODUCTION

The literature review will discuss the various methods and approaches used in the prediction and detection of financial distress in various organisations. Both classical statistical methods such as MDA, Logit and Probit and machine learning methods such as Bagging, Boosting and Ensemble Methods have proved effective but they can always be improved upon with the aim of creating better prediction models.

Over the course of running an organisation, various decisions have to be made to ensure that profit is maximised as they meet their monthly or yearly financial obligations as stated by Ikpesu & Eboiyehi(2018). If companies fail to service their financial obligations by the due date and time, this could affect the organisation due to loss of financial income, loss of investor trust and the community through loss of jobs which inadvertently results in people suffering from depression among other health problems (Valenzuela et al., 2022).

2.2 EXISTING MEASURES FOR FINANCIAL DISTRESS PREDICTION

2.2.1 STATISTICAL APPROACHES

Valaskova et al., (2020) conducted a study on predicting financial distress on 3329 Slovak enterprises operating in the agricultural sector located in Slovakia. The dataset was sourced from the NACE A sector and it contained the financial reports from 2016,2017 and 2018. The three-year long period was chosen as a response to a newly adopted legislation that established the status of a company in a financial crisis. The new Act No. 7/2005 Coll. on bankruptcy and restructuring and Act No 513/1991 Coll. of the Commercial Code states that an enterprise is considered to be in crisis when it's in decline or in a threat of bankruptcy. If a company's equity-to-liability ratio is less than 8 per 100, then it is at risk of bankruptcy. Their models such as the Altman Model, Agricultural model of SR and general Slovak models were based on multiple discriminant analysis (MDA). Area Under the Curve (AUC) was used to gauge the Agricultural

model of SR's accuracy which averaged the value of 0.863. A sensitivity of 62.11% and specificity of 87.39% were also achieved.

Noga and Adamowicz (2020) used the MDA research method during the development of their forecasting bankruptcy of wood enterprises (FMWE) model. The FMWE model was constructed and tested using financial data from 135 wood industry enterprises, whose selections were on the basis of the Polish Classification of Activities (the Code List of Classification of Business Activities in Poland PKD). The financial data were obtained from District Courts and credit information bureaus. Using the FMWE model, an AUC of 89% was achieved after using five financial indicators; (i) current assets or liabilities, (ii) total income/average total assets; (iii) private capital/ total debt, (iv) profit from the operating activity—depreciation, product sales, (v) operating cost/ short-term commitments.

Svabova et al., (2022) created three prediction models for small and medium-sized companies in Slovakia, based on real data that was sourced from the Amadeus database from 2016-2018. The financial indicators of the selected companies were measured in 2016 and 2017 and the result variables, indicating the prosperity or non-prosperity of the company, were set in 2017 and 2018. The whole database contained (after dataset preparation including the correctness check, missing data analysis and analysis of outliers) 75,652 companies. The three models, 1-year prediction model using the data from 2016 (and prosperity of the company from 2017); 1-year prediction model using the data from 2017 (and prosperity of the company from 2018) and 2-year prediction model using the data from 2016 (and prosperity of the company from 2018) were based on the combination of two research methods, discriminant analysis and logistic regression. All models achieved an AUC of 0.806,0.890 and 0.856 respectively.

Mohammed and Moudden (2020) used logistic regression to determine the reasons for the financial failures of SMEs. They retrieved the data of healthy and failed companies from the Moroccan Bank. Their findings state that the Autonomy ratio, interest to sales, asset turnover, days in accounts receivable, and duration of trade payables increase the probability of financial failure, while repayment capacity and return on assets reduce the probability of failure. On the other hand, given variables show an overall classification rate of healthy and failing SMEs of 91.11% three years before failure and 84.44% two years and one year before failure.

2.2.2 MACHINE LEARNING APPROACHES

Hamori et al., (2018) used three ensemble-learning methods—, bagging, random forest, and boosting—and various neural-network methods, each of which had a different activation functions such as Tanh, ReLU, TanH with Dropout and ReLU with Dropout to assess credit risk in financial institutions whose core business is lending. The dataset used was sourced from UCI machine learning repository which had a total number of 30,000 observations. Boosting and Random Forest methods had AUC values of 76.9% and 60.5% and F-scores of 74.4 % and 71.4% respectively, which outperformed NN and DNN that had AUC scores of 70.4% and 75.1% respectively.

Barboza et al., (2018) conducted a study on bankruptcy prediction using SVM, bagging, boosting and random forest models using financial data on American and Canadian companies from 1985 to 2013 using Compustat. A training set that included information on 449 companies that filed for bankruptcy during that period without the use of normalisation. Insolvent firms in their training set included all companies in the database that went bankrupt and whose data was available three years prior to their bankruptcy filing. Bagging, Boosting and random forest achieved Accuracy values of 86.65%,85.67% and 87.06% respectively.

Pawelek (2017) used financial data of companies operating in Poland's industrial processing sector in the years 2005-2008. Data classification methods such as: a classification tree, KNN algorithm, SVM, a neural network, RF, bagging, boosting, naive Bayes, logistic regression and LDA were employed. The predictive accuracy of the constructed models was assessed using sensitivity and specificity and calculations were done using R programming language. Overall, a sensitivity of 74.2% and specificity of 93.4 % was achieved.

Xu et al., (2019) predicted financial distress in Chinese financial institutions using data from more than 150 companies. The data set was sourced from the Baidu website in the seven-year period from 2011-2017. Stepwise LR was applied to identify six optimal traditional variables: working capital/total asset, current debt/sales, retained earnings/total asset, cash flow/sales, market value equity/total debt, and net income/total asset from 18 popular financial variables. A novel soft ensemble model (ANSEM) was proposed and it had superior performance in comparison to ES (expert system method), CNN, EMEW (ensemble model based on equal weight), EMNN (ensemble model based on the convolutional neural network), and EMRD

(ensemble model based on the rough set theory and evidence theory) as it had an accuracy of 80% after employing the use of 5-fold cross-validation.

Liang et al., (2020) applied the stacking ensemble model to predict bankruptcy in US companies based on 40 financial ratios and 21 corporate governance indicators from the UCLA-LoPucki Bankruptcy Research Dataset between 1996 and 2014, which contained financial data from 764 bankrupt firms available from the database. Stacking ensembles are based on a two-level architecture, the first level constructs a number of different classifiers (i.e. base learners) whose outputs are used to train the second level classifier (i.e. meta learner) for the final prediction result. A mean accuracy of 89.8% was achieved.

Ekini and Erdal (2017) predicted bank failures in Turkey based on 35 financial ratios using conventional machine learning models, ensemble learning models and hybrid ensemble learning models. The ratios are divided into six groups: capital ratios, assets quality, liquidity, profitability, income–expenditure structure, and activity ratios (CAMELS). The hybrid ensemble learning model is better than conventional machine learning models in classification accuracy, sensitivity, specificity, and ROC curves with scores of 83.783%, 0.838, 0.164 and 0.89 respectively.

Kim and Upneja (2021) developed business failure prediction models based on American restaurants between 1980 and 1970 using the majority voting ensemble method with a decision tree. The period is divided into economic recession and economic growth. The three models: models for the entire period, models for the economic downturn, and models for economic expansion based on data from the Compustat database by Standard and Poor's Institutional Market Services in WEKA 3.9. The sample includes 2747 observations, 1432 of which failed. The selected models achieved different accuracies: the model for the entire period (88.02%), the model for an economic downturn (80.81%), and the model for economic expansion (87.02%).

Tran et al., (2022) predicted financial distress using the financial statements of Vietnamese companies listed on the Ho Chi Minh Stock Exchange, Hanoi Stock Exchange, and UPCOM. Data were collected from 2010 to 2021. The companies were chosen if their operational income was negative for three consecutive years. The data had 3277 observations, of which 436 companies were in financial distress, the SMOTE technique was used to handle imbalances in the data. Extreme Gradient Boosting and Random Forest achieved an F1 score of 0.85 and 0.84.

2.3 CHALLENGES FACED BY EXISTING APPROACHES

AUTHOR	DATASET USED	ALGORITHMS USED	BENEFITS	CHALLENGES
Valaskova et al., (2020)	The dataset was sourced from the NACE A sector and it contained the financial reports from 2016,2017 and 2018	MDA.	An accuracy of 0.863. A sensitivity of 62.11% and specificity of 87.39% were achieved	No challenges were encountered.
Noga and Adamowicz (2020)	Financial data from 135 wood industry enterprises.	MDA	An AUC of 89% was achieved	Small dataset with small amount of variables
Svabova et al., (2022)	A dataset from the Amadeus with data from 75,652 companies.	Discriminant analysis and logistic regression	An AUC of 0.890 was achieved.	Dataset had a lot of missing values and outliers.
Mohammed and Moudden (2020)	A dataset containing failed companies from the Morroccan Bank.	Logistic Regression	Achieved an accuracy of 91%.	Dataset was highly Imbalanced.
Hamori et al., (2018)	The dataset was sourced from UCI machine learning repository which had a total	bagging, random forest, and boosting	Boosting and Random Forest methods had AUC values of 76.9% and 60.5% and F-	Dataset contained a lot of outliers and was highly Imbalanced. No dimensionality reduction

	number of 30,000 observations		scores of 74.4 % and 71.4% respectively	techniques were used.
Barboza et al., (2018)	Financial data on American and Canadian companies from 1985 to 2013 was used.	SVM, bagging, boosting and random forest	Bagging, Boosting and random forest achieved Accuracy values of 86.65%,85.67% and 87.06% respectively	The Data used was insufficient.
Pawelek (2017)	financial data of companies operating in Poland's industrial processing sector in the years 2005-2008	KNN algorithm, SVM, a neural network, RF, bagging, boosting, naive Bayes, logistic regression and LDA	Overall, a sensitivity of 74.2% and specificity of 93.4 % was achieved.	No challenges were reported.
Xu et al., (2019)	Data about 150 Chinese financial Institutions was sourced from the Baidu website.	Stepwise LR, CNN	An accuracy of 80% after employing the use of 5-fold cross-validation was achieved	Dataset used was prone to noise
Liang et al., (2020)	The UCLA-LoPucki Bankruptcy Research Dataset	Stacking Ensemble methods	A mean accuracy of 89.8% was achieved.	Dataset contained missing values and a lot of outliers.

	between 1996 and 2014			
Ekini and Erdal (2017)	A dataset containing failed banks in Turkey.	Machine learning models, ensemble learning models and hybrid ensemble learning models	The hybrid ensemble learning model had classification accuracy, sensitivity, specificity, and ROC curves values with scores of 83.783%, 0.838, 0.164 and 0.89 respectively	No challenges were reported.
Kim and Upneja (2021)	Data on American restaurants between 1980 and 1970 with 2747 observations	the majority voting ensemble method with a decision tree.	An Overall accuracy of 84% was achieved.	Dataset contained missing values and outliers.
Tran et al., (2022)	the financial statements of Vietnamese companies listed on the Ho Chi Minh Stock Exchange, Hanoi Stock Exchange,	Extreme Gradient Boosting and Random Forest	XGB and RF achieved an F1 score of 0.85 and 0.84	Dataset was highly unbalanced.

	and UPCOM. Data were collected from 2010 to 2021			

Table 2.3.1 Challenges Faced by existing approaches.

CHAPTER THREE

3.0 METHODOLOGY

3.1 INTRODUCTION

This chapter describes the steps that will be undertaken to design the algorithms and models for financial distress prediction. The Data that will be collected is secondary as they were obtained from articles already written by someone. Articles and journal reviewing was done and the tools used were websites such as Google Scholar, Hindawi, Research Gate, Wikipedia among others. The algorithms are to be written in python and run on Jupyter IDE and Google Collab. Other libraries designed specifically for machine learning research will also be used.

The following data flow diagram shows the steps followed for building the model.

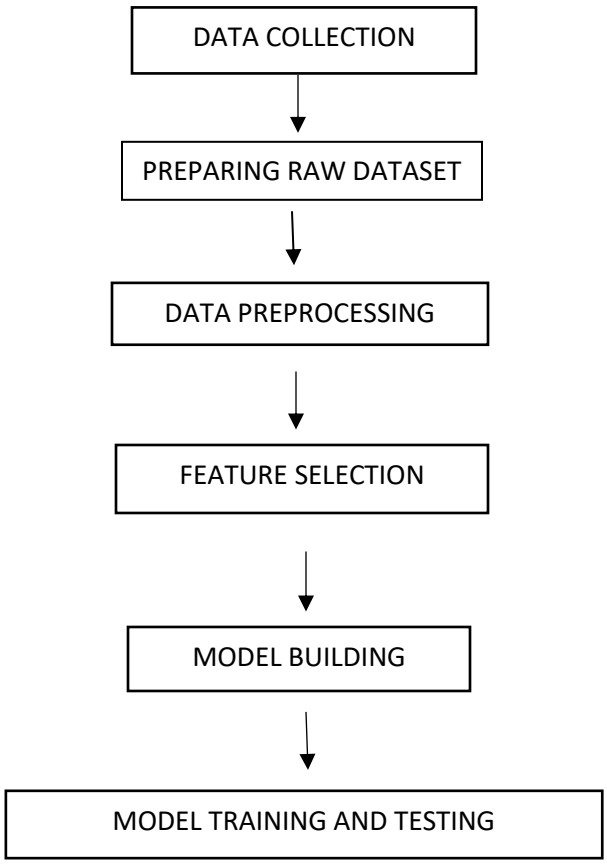


Figure 3.1.1 Flowchart Diagram

3.2 DATASET DESCRIPTION

The dataset used for this study was obtained from Kaggle. This dataset provides real-world indicators of Brazilian enterprises based on data from the Brazilian Securities and Exchange Commission, organized quarterly in a non-stationary manner. The dataset covers 10 years (2011 to 2020) and includes 905 different corporations together with their financial and non-financial characteristics and 23,834 records, each described with 84 indicators.

The indicators are shown in the table below.

Attribute information	
Column	**Description**
ID	Sequential value for different enterprises
QUARTER	The last day of the quarter e.g., '2011-03-31', '2011-06-30', '2011-09-30' and '2011-12-31' for the first, second, third and fourth quarters, respectively.
A1	Total assets
A2	Current assets
A3	Availability
A4	Receivables
A5	Inventory
A6	Long-term assets
A7	Intangible assets
A8	Tangible assets
A9	Fixed assets
A10	Accumulated depreciation
A11	Accumulated amortization
A12	Investments
A13	Total liabilities
A14	Current liabilities
A15	Non-current liabilities
A16	Commitments (A13 - A14)
A17	Net worth (A12 - A15)

A18	Share capital
A19	Reserves (revenue reserves + capital reserves + non-cash reserve)
A20	Provisions
A21	Long term loan
A22	Gross income
A23	Expenses
A24	Net earnings
A25	Operating expenses
A26	Operating profit
A27	Financial result
A28	Financial expenses
A29	Profit before tax
A30	Tax expenses
A31	Net income
A32	Cash flows from operating activities
A33	Cash Flows from Investing
A34	Cash Flows from Financing
A35	Outstanding shares
A36	Current ratio
A37	Quick ratio
A38	Cash ratio
A39	Interest coverage ratio
A40	Debt ratio
A41	Tangible asset coverage ratio
A42	Ratio of equity to debt
A43	Ratio of commitments to tangible assets
A44	Liquidity ratio
A45	Receivable assets ratio
A46	Fixed Asset Ratio (FAR)
A47	Ratio of stockholders' equity to fixed assets
A48	Current debt ratio
A49	Operating net profit ratio
A50	Ratio of receivables to gross income

A51	Ratio of inventory to income
A52	Inventory turnover
A53	Turnover ratio of account payable
A54	Turnover of current assets
A55	Ratio of fixed assets to income
A56	Total capital turnover
A57	Return On Assets (ROA)
A58	Ratio of net profit to total assets
A59	Ratio of net profit to current assets
A60	Ratio of net profit fixed assets
A61	Return On Equity (ROE)
A62	Operating profit ratio
A63	Ratio of total operating cost to gross revenue
A64	Expenses to sales Ratio (ER)
A65	Management Expense Ratio (MER)
A66	Financial Expense Ratio (FER)
A67	Free Cash Flow (FCF)
A68	Ratio of operating cash to net profit
A69	Ratio of operating cash to income
A70	Cash recovery rate
A71	Financial leverage
A72	Operational leverage
A73	Combined leverage
A74	Growth of capital maintenance rate
A75	Growth of capital accumulation rate
A76	Growth of total assets rate
A77	Growth rate of ROE
A78	Growth rate of net profit
A79	Growth rate of operating profit
A80	Growth rate of operating receipt
A81	Growth rate of operating cost
A82	Earnings per share
A83	Net asset value per share (NAVPS)

A84	Net cash per share

Table 3.2.1 Dataset attributes.

3.3 THEORETICAL DESCRIPTION OF CLASSIFIERS

3.3.1 DECISION TREES

Decision Trees are constructed using two kinds of elements: nodes and branches. At each node, one of the features of our data is evaluated in order to split the observations in the training process or to make a specific data point follow a certain path when making a prediction. They are paths where different variables are evaluated that lead to a leaf node where similar observations are grouped as stated by Faria et al., (2021).

In the training process the tree is built by analysing the possible features along with their values and deciding which features best split our data so that different data points go to one side of the split while minimizing some kind of error. They are very simple and sometimes tend to overfit (learn the training data exceptionally well and have a hard time generalizing to new data points). However, despite these drawbacks, they are probably the most intuitive and easy Machine Learning model to understand, and by examining which path our data points follow, we can easily know why a tree made a certain prediction.

3.3.2 RANDOM FOREST

It is can be used both for regression and classification. They are one of the most popular ensemble methods that involve using many learners to enhance the performance of any single one of them individually. Random Forests are collection of many individual Decision Trees and introduce randomness and numbers into the equation, fixing many of the problems of individual decision trees, like overfitting and poor prediction power.

In Random Forest, each tree is built using a subset of the training data, and a subset of the possible features. As more and more trees are built, a wider range of data is used, and more features come into play, making very strong, aggregated models. Individual trees are built

independently, using the same procedure as for a normal decision tree but with only a random portion of the data and only considering a random subset of the features at each node. Aside from this, the training procedure is exactly the same as for an individual Decision Tree, repeated N times thus making this algorithm simple, flexible and fast (Faria et al., 2021).

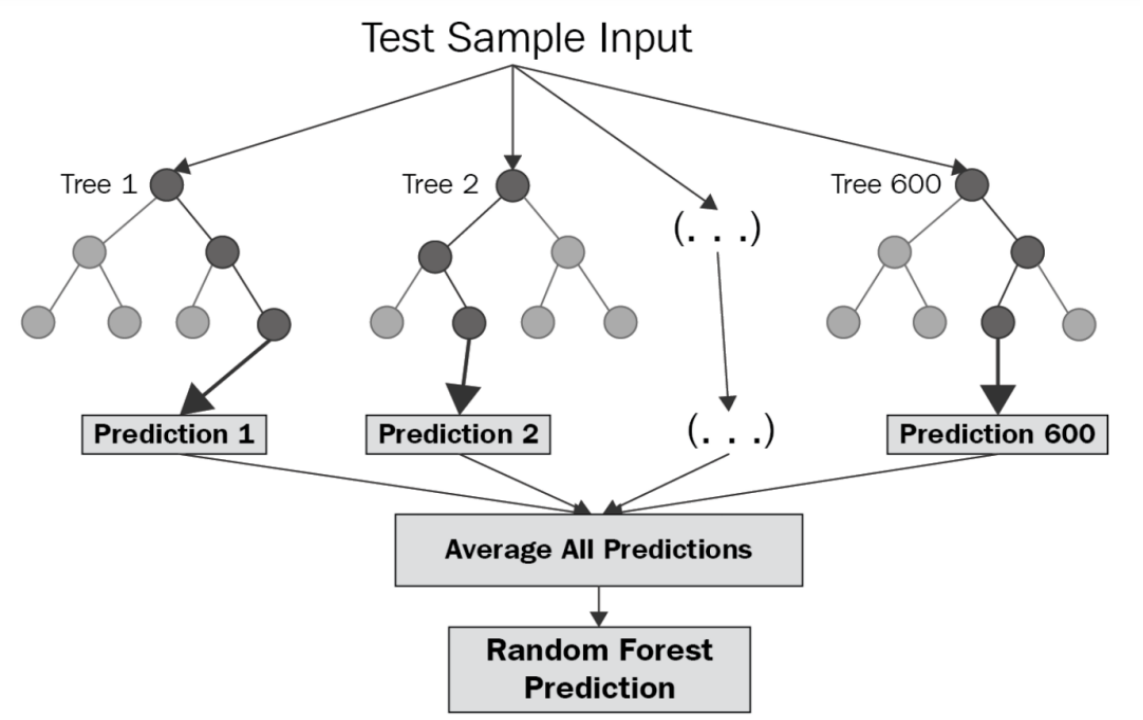


Figure 3.3.2.1 How Random Forest and Decision Trees work

3.3.3 SUPPORT VECTOR MACHINES

SVM is a supervised learning algorithm that is commonly used for classification and regression problems. SVM works by finding the hyperplane that maximally separates the two classes (healthy and diseased), and classifying new data points based on which side of the hyperplane they fall on.

The advantage of using SVM in this model is its ability to handle non-linearly separable data and handle high-dimensional data. Also, SVM can balance the search for complex models and

learning ability even without sufficient large sample data (Altan et. al., 2022). SVM also has the ability to handle imbalanced datasets, which is common in the majority of financial distress datasets, where the number of financially healthy companies may far outweigh the number of financially unhealthy companies.

In conclusion, the use of SVM in the proposed financial distress detection system is motivated by its ability to handle complex data, handle imbalanced datasets, and provide accurate and reliable results. The use of SVM in this model has the potential to improve the accuracy and reliability of financial distress prediction, making it a valuable tool for investors and large companies.

3.4 DATA PRE-PROCESSING

Data preprocessing is a critical step in the data preparation pipeline, laying the foundation for effective machine learning models. It encompasses a series of techniques and operations that aim to clean, transform, and refine raw data into a format that will make it suitable for analysis and modeling. The first initial task will involve handling missing values by using mean imputation and regression imputation to estimate values based on relationships with other variables. Additionally, outliers, which can skew the learning process of a model, will be identified and treated appropriately. This will be achieved through winsorization by limiting extreme values in the dataset by capping them at a certain threshold, or by removing them entirely.

Feature scaling, will ensure that all variables contribute equally to the learning process. Common methods include standardization (scaling to a mean of 0 and standard deviation of 1) or min-max scaling (scaling to a specified range, often [0, 1]). Categorical variables, if any, which are non-numeric in nature, will also be encoded into a numerical format for the model to process effectively through techniques like one-hot encoding.

3.5 EVALUATION METRICS

a. F1 Score

F1 score is a widely used performance metric to evaluate the performance of binary classification models. It is the harmonic mean of precision and recall. Precision is the fraction of correctly classified positive instances among all instances classified as positive, while recall is the fraction of correctly classified positive instances among all actual positive instances. The mathematical formula for the F1 score is:

F1 Score = 2 * (precision * recall) / (precision + recall)

b. Accuracy

Accuracy is a metric used to evaluate the performance of classification models. It measures the proportion of correct predictions made by the model over the total number of predictions. The mathematical formula for calculating accuracy is:

Accuracy = (Number of correct predictions) / (Total number of predictions)

c. Recall Score

Recall is a performance metric in machine learning that measures the ability of a model to identify all relevant instances (True Positive) while minimizing the number of false negatives (False Negative). In other words, it measures the ability of a model to find all the relevant data points of a particular class. The mathematical expression of recall is given by:

Recall = True Positive / (True Positive + False Negative)

d. Precision

This metric is used for the determination of the proportion of positive prediction that was actually correct. It is mathematical formula is;

Precision=$\frac{tue\ positive}{(TP+FP)}$

e. Confusion Matrix

A confusion matrix is a tabular representation of prediction outcomes of any binary classifier, which is used to describe the performance of the classification model on a set of test data when true values are known. The binary classifier has only two values, true which is indicated by 1 and false, which is indicated by 0. A typical confusion matrix for a binary classifier looks like the below image (However, it can be extended to use for classifiers with more than two classes).

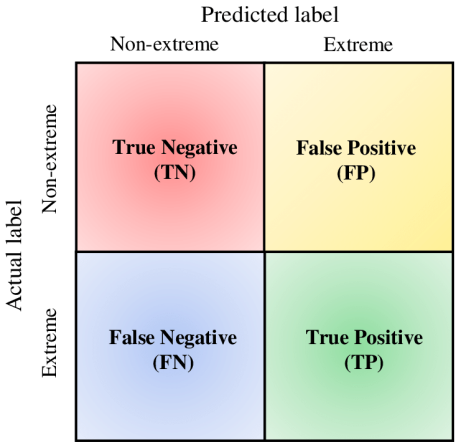


Figure 3.5.1 How Confusion Matrix Works

The table has four terminologies, which are as follows:

1. True Positive (TP): In this case, the prediction outcome is true, and it is true in reality, also.
2. True Negative (TN): in this case, the prediction outcome is false, and it is false in reality, also.
3. False Positive (FP): In this case, prediction outcomes are true, but they are false in actuality.
4. False Negative (FN): In this case, predictions are false, and they are true in actuality.

3.6 PSEUDOCODE

3.6.1 Decision Trees

1. Place the best attribute of the dataset at the root of the tree.
2. Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

3.6.2 Random Forest

1. Randomly select “k” features from total “m” features.
    1. Where k << m
2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until “l” number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.
6. Takes the **test features** and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
7. Calculate the **votes** for each predicted target.
8. Consider the **high voted** predicted target as the **final prediction** from the random forest algorithm.

3.7 FEATURE SELECTION

Feature selection is performed so as to obtain features that would result in making the model perform better depending on the chosen performance metrics resulting into higher levels of prediction and detection of said problem. The dataset contained many features out of which we chose the ones that were most important. Hence, having only the significant attributes and reducing the number of irrelevant attributes can increase the performance of classifiers.

This will be performed using PCA. Wang et. al.,(2017) states that the basic principle of PCA is that when there are too many variables, it will mine and find the correlation between variables and the subject, delete redundant related variables, and create new variables as few as possible according to the overlap between variables, so as to make the correlation between these variables as low as possible or no correlation, and minimize the number of variables without affecting the accuracy of the main research results and reducing the loss of information.

The new variable obtained by PCA is called principal component. When the observed value is m and the attribute value is p, the number of principal components is min ($m - 1$, p). If there are n row data and a characteristic dimension in the research object, x_{ij} is the i-dimension attribute in the j row data, and there is matrix X at this time. Matrix X is a $n \times a$ matrix, and its covariance matrix is shown as:

$$C = \frac{1}{n} X^{t+1} \qquad (1)$$

In (1), t is the number of time series. The covariance matrix C is a symmetric matrix of $a \times a$, and its diagonal is the variance of each eigenvalue. Covariance matrix C is also a real symmetric matrix, which has some properties of real symmetric matrix, and these properties can be used to obtain several nonzero eigenvectors to form a new matrix. From matrix E, a new matrix Λ can be obtained, as follows:

$$\Lambda = Et+1C. \qquad (2)$$

In order to reduce the redundant data, the data in the characteristic matrix X are converted to another characteristic space to obtain a new matrix Y. The eigenvector e_i corresponding to each eigenvalue needs to satisfy equation (3). In equation (3), i 1, 2, ..., p and e_{ij} are j vectors of e_i

$$\sum_{j=1}^{p} e_{ij}^{2} = 1. \quad (3)$$

Since each characteristic of matrix *Y* is expected to be linear independent, its covariance matrix *D* is also a diagonal matrix, its diagonal is variance, that is, the blank is covariance. When the covariance is 0, it means that the two vectors are orthogonal. The eigenvectors of the matrix *C* form a new matrix *Z*, and the covariance matrix *D* is shown as

$$D = Zt+1C=\frac{1}{n}Yt+1. \quad (4)$$

Each value on the diagonal of matrix *D* is the eigenvalue of covariance matrix *C* (Sinclair et. al., 2022). The eigenvalues and eigenvectors in the matrix *D* are sorted from small to large and from left to right, respectively, and then the first *k* are taken for compression conversion to obtain the dimension reduced data matrix *Y*. In matrix *X*, there is no correlation between the principal components of the linear combination of its attributes, and the principal components have the maximum variance in the linear combination of various attributes. The proportion of total variance from the first to *n* the second principal component decreases in turn. In general, several principal components with a cumulative proportion of more than 85% can be considered as less information loss, which can be adopted as said by Basher and Hallam (2022). The flowchart of the whole process is shown below;

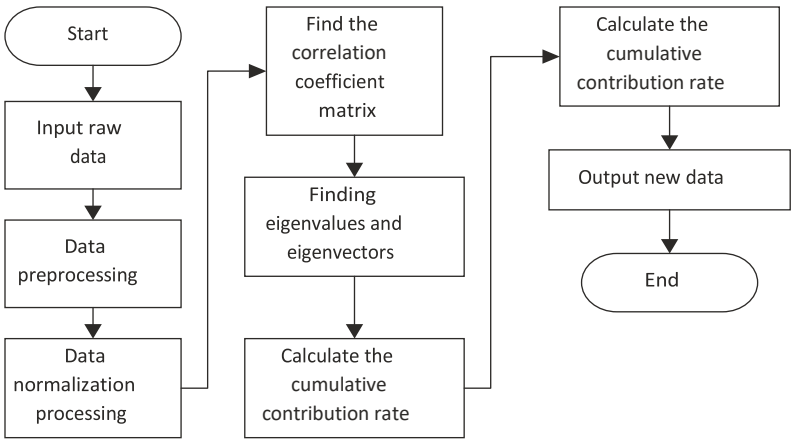


Figure 3.7.1 PCA flowchart

3.8 EXPERIMENTAL SET UP AND DESIGN

The algorithms and models will be built and deployed in windows and Linux operating systems using python and its libraries for machine learning. Data preprocessing, training and testing will be performed and Jupyter IDE will be used for building and evaluation of the model using python. After selecting the necessary features and removing noise, the dataset will be split into training and testing data that will be used to train and test the model built using random forest and decision trees classifiers. The model will then be integrated to a web app built using JavaScript's React framework and communication between the model and website will be via FAST API.

REFERENCES

Nathan Bomey. (2020). Brooks Brothers store closings planned: Retailer files for Chapter 11 bankruptcy protection. *https://www.usatoday.com/story/money/2020/07/08/brooks-brothers-store-closings-brooks-brothers-files-chapter-11-bankruptcy/5396613002/*

Bhasin, H. (2021, June 29). Claire's Stores files for bankruptcy, closing up to 200 stores. CNBC. *https://www.today.com/style/claire-s-will-reportedly-file-bankruptcy-t124786*

Chris Casey (2023). Aero-Farms files for Chapter 11 bankruptcy protection. *https://www.fooddive.com/news/aerofarms-files-chapter-11-bankruptcy-protection/652598/*

Borrescio-Higa F, Droller F and Valenzuela P (2022) Financial Distress and Psychological Well-Being During the COVID-19 Pandemic.
Int J Public Health 67:1604591.doi: 10.3389/ijph.2022.1604591

Ingrid Lexova and Umer Khan (2023). July filings propel 2023 US corporate bankruptcy tally past 2022's total.
https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/july-filings-propel-2023-us-corporate-bankruptcy-tally-past-2022-s-total-76838356

Ikpesu, F., & Eboiyehi, O. C. (2018). Capital structure and corporate financial distress of manufacturing firms in Nigeria. Journal Of Accounting and Taxation.
http://www.academicjournals.org/JAT

American Psychological Association. (2018). Stress in America: Financial stress and mental health. *https://www.apa.org/news/press/releases/stress/2014/financial-stress*

Desheng Wu , Xiyuan Ma , David L. Olson (2022).Financial distress prediction using integrated Z-score and multilayer perceptron neural networks. Retrieved from *https://doi.org/10.1016/j.dss.2022.113814*

Centre for American Progress. (2021). The financial security of single parents. *https://journals.sagepub.com/doi/full/10.1177/00027162221122682*

Inekwe, J.; Jin, Y.; Valenzuela, M.R. The effects of financial distress: Evidence from US GDP growth. Econ. Model. 2018. Retrieved from *https://www.sciencedirect.com/science/article/abs/pii/S0264999317315791?via%3Dihub*

Economic Policy Institute. (2020). Financial distress among older adults. *https://bmcgeriatr.biomedcentral.com/articles/10.1186/s12877-020-01687-5*

S. Hamori, M. Kawai, T. Kume, Y. Murakami, C. Watanabe.(2018). Ensemble learning or deep learning? Application to default risk analysis. *https://www.mdpi.com/1911-8074/11/1/12*

Consumer Financial Protection Bureau. (2022). Financial hardship and access to essential services. *https://www.consumerfinance.gov/consumer-tools/educator-tools/resources-for-older-adults/*

Flavio Barboza, Herbert Kimura , Edward Altmanc. (2018) Machine learning models and bankruptcy prediction. *https://sci-hub.se/10.1016/j.eswa.2017.04.006*

Pawełek, B (2017). Prediction of Company Bankruptcy in the Context of Changes. *In Proceedings of the 11th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena: Conference Proceedings, Zakopane, Poland, pp. 290–299.*

Xu, W.; Fu, H.; Pan, Y(2019). A Novel Soft Ensemble Model for Financial Distress Prediction with Different Sample Sizes. *https://www.hindawi.com/journals/mpe/2019/3085247/*

Liang, D.; Tsai, C.-F.; Lu, H.-Y. (Richard); Chang, L.-S (2020). Combining Corporate Governance Indicators with Stacking Ensembles for Financial Distress Prediction. J. Bus. Res. 2020, 120, 137–146.
https://www.semanticscholar.org/paper/Combining-corporate-governance-indicators-with-for-Liang-Tsai/7fad075484162b0ba5fc383b00342a4bc942148f

Aykut Ekinci & Halil İbrahim Erdal, 2017. Forecasting Bank Failure: Base Learners, Ensembles and Hybrid Ensembles, *Computational Economics, Springer; Society for Computational Economics, vol. 49(4), pages 677-686, April.*

Kim, S.Y.; Upneja (2021). A. Majority Voting Ensemble with a Decision Trees for Business Failure Prediction during Economic Downturns.
https://www.sciencedirect.com/science/article/pii/S2444569X21000081?via%3Dihub

Zizi, Y. - Mohamed, O. - Moudden, A. (2020): Determinants and Predictors of SMEs' Financial Failure: A Logistic Regression Approach. Risks.
8. DOI: 10.3390/risks8040107

American Psychological Association. (2023). Mental health and racial and ethnic minority communities. *https://www.apa.org/pi/families/resources/mental-health-needs.pdf*

Q. Ding, Z. Shao, X. Huang, O. Altan, and Y Fan, “Improving urban land cover mapping with the fusion of optical and SAR data based on feature selection strategy,” *Photogrammetric Engineering & Remote Sensing, vol. 88, no. 1, pp. 17–28, 2022.*

Faria Ferdowsy, Kazi Samsul Alam Rahi, Md. Ismail Jabiullah, Md. Tarek Habib. (2021). A machine learning approach for obesity risk prediction
https://www.researchgate.net/publication/353713038_A_Machine_Learning_Approach_for_Obesity_Risk_Prediction

M. Sun, Y. Wang, and J. Liu (2017). Generalized Peaceman-Rachford splitting method for multiple-block separable convex programming with applications to robust PCA. *Calcolo, vol. 54, no. 1, pp. 77–94.*

A. R. M. A. Basher and S. J. Hallam (2022), "Aggregating statistically correlated metabolic pathways into groups to improve prediction performance," *The 15th International Joint Conference on Biomedical Engineering Systems and Technologies, pp. 49–61, Singapore, January2022.*

K. Batpurev, S. J. Sinclair, O. Avirmed, M. P. Scroggie, K. Olson, and M. D White (2022) "Stakeholders from diverse backgrounds make similar judgments about ecological condition and collapse in Mongolian rangelands," *Conservation Science and Practice, vol. 4, no. 1, p. e574.*

National Bureau of Economic Research. (2022). Financial distress among people with disabilities. *https://www.aeaweb.org/articles?id=10.1257/app.20190709*