UNIVERSITY OF PARDUBICE
FACULTY OF ECONOMICS AND ADMINISTRATION
INSTITUTE OF SYSTEM ENGINEERING AND INFORMATICS

FORECASTING REGIONAL FINANCIAL PERFORMANCE USING SOFT-COMPUTING METHODS

DISSERTATION THESIS

Author: Ing. Evelyn Toseafa
Supervisor: prof. Ing. Petr Hájek, Ph.D.

Pardubice 2023

AUTHOR'S DECLERATION

I hereby declare:

This thesis entitled "Forecasting Regional Financial Performance Using Soft-Computing Methods" is my own work. All literary sources and information that I used in the thesis are referenced in the bibliography.

I have been acquainted with the fact that my work is subject to the rights and obligations arising from Act No. 121/2000 Sb., On Copyright, on Rights Related to Copyright and on Amendments to Certain Acts (Copyright Act), as amended, especially with the fact that the University of Pardubice has the right to conclude a license agreement for the use of this thesis as a school work under Section 60, Subsection 1 of the Copyright Act, and that if this thesis is used by me or a license to use it is granted to another entity, the University of Pardubice is entitled to request a reasonable fee from me to cover the costs incurred for the creation of the work, depending on the circumstances up to their actual amount.

I acknowledge that in accordance with Section 47b of Act No. 111/1998 Sb., On Higher Education Institutions and on Amendments to Other Acts (Act on Higher Education Institutions), as amended, and the Directive of the University of Pardubice No. 7/2019 Rules for Submission, Publication and Layout of Theses, as amended, the thesis will be published through the Digital Library of the University of Pardubice.

In Pardubice on 30. 12. 2023 Ing. Evelyn Toseafa

Acknowledgements

Firstly, I would like to express my deepest gratitude to Professor Petr Hájek for his invaluable contribution to this work. His guidance, knowledge, and resources have been instrumental in this research. I aspire to embody his exemplary character and scholarly qualities.

I am also very grateful to Professor Petr Hájek for his continuous supervisory support and counsel throughout my doctoral research. I also appreciate Mrs. Bartonickova Radka for her support and cooperative spirit during my stay at the University.

I would like to express my deepest gratitude to my siblings, Pickinson and Chrysler Adams, for their unconditional love and unwavering support before and during my academic years.

I would like to acknowledge and give my warmest gratitude to my cousin Matilda Efua Nufouba Hammond, for her unconditional love and support during her stay here with me in the Czech Republic. I am forever grateful.

I am extremely grateful to Associate Professor James Beguah Odei of the Ohio State University, USA, Dr. Samuel Amponsah Odei of the University of Hradec Kralove, Ms. Debbie Barnie of The World Bank Washington DC. USA, and Dr. Michael Karikari Appiah for their invaluable guidance, encouragement, and motivation throughout my studies at this university. Their contributions were crucial to the fruition of this research.

The basis of my identity is undoubtedly attributed to my parents, Mr. W.T. D. Essilfie and the late Mrs. Dorothy Essilfie, as well as Late Jonathan Addo, Mr. Paul Acquah Wilson, and Mr. Edison Amanor for their unwavering love, support, and guidance. Without their sacrifices, I would not have reached this pinnacle of my academic journey.

I am also very grateful to my ex-husband, Dr. Hero Kwabena Toseafa, for the role he played in bringing me to the Czech Republic to reach this academic milestone despite all odds.

To my loving and supportive husband, Mr. Michael Kweku Awuah Adu-Dankwa, whose steadfast love, encouragement, and advice has spear headed me to a successful conclusion of this study.

I would also like to thank Godfred Enim Douglas and my best friends for their staunch support.

Finally, I would like to express my gratitude to the University of Pardubice for funding my studies, which has been instrumental in achieving this significant academic milestone.

ANNOTATION

The difficulty in resolving the issues associated with forecasting regional financial performance has spurred the emergence of various applications of soft computing methods to tackle these challenges. This has inspired the development of hybrid models that employ diverse soft computing techniques. In this work, different machine learning methods such as random forest, XGBoost, support vector machines, neural networks, and fuzzy rule-based systems are utilized to improve the prediction of regional financial performance. I propose a novel hybrid method that integrates feature selection, class balancing, and ensemble classifiers in a cost-sensitive prediction scenario. More precisely, the proposed approach aims to develop an accurate decision support system that minimizes the misclassification cost in credit rating classification for sub-sovereign entities across various countries and world regions. Cost-sensitive learning is employed to adjust the training instances in accordance with the total cost associated with each class, facilitating the prediction of nominal rating classes at a lower misclassification cost. Furthermore, it is demonstrated that combining bagging with decision trees as base learners can mitigate the risk of overfitting, a common issue in individual machine learning methods. To validate the proposed approach, I have conducted experiments using two different types of datasets from Moody's credit rating agency. The results show that the proposed hybrid model surpasses existing forecasting models in terms of misclassification cost and other classification metrics.

Keywords: regional financial performance, sub-sovereign, credit rating, soft-computing method, machine learning, ensemble methods, cost sensitive learning

NÁZEV

Predikce regionální finanční výkonnosti pomocí soft-computingových metod

ANOTACE

Potíže s řešením problémů spojených s predikcí regionální finanční výkonnosti podnítily vznik různých metod soft computingu. To inspirovalo vývoj hybridních modelů, které kombinují různé soft computingové techniky. V této práci se ke zlepšení predikce regionální finanční výkonnosti využívají různé metody strojového učení, jako jsou náhodné lesy, XGBoost, podpůrné vektorové stroje, neuronové sítě a systémy založené na fuzzy pravidlech. Navrhuji nový hybridní model, který integruje selekci proměnných, vyvažování tříd a soubory klasifikátorů v nákladově citlivém predikčním scénáři. Přesněji řečeno, navrhovaný přístup si klade za cíl vyvinout přesný systém podpory rozhodování, který minimalizuje náklady na nesprávnou klasifikaci při predikci úvěrového ratingu pro regionální subjekty v různých zemích. Nákladově citlivé učení se používá k úpravě vah instancí při učení v souladu s celkovými náklady spojenými s každou třídou, což usnadňuje predikci nominálních ratingových tříd při nižších nákladech na nesprávnou klasifikaci. Kromě toho je prokázáno, že kombinování baggingu s rozhodovacími stromy jako základními klasifikátory může zmírnit riziko přeučení, což je běžný problém u jednotlivých modelů strojového učení. Pro ověření navrhovaného přístupu byly provedeny experimenty s použitím dvou různých typů datových souborů od ratingové agentury Moody's. Výsledky ukazují, že navrhovaný hybridní model předčí stávající predikční modely, zejména pokud jde o náklady na nesprávnou klasifikaci a další klasifikační metriky.

Klíčová slova: regionální finanční výkonnost, region, úvěrový rating, metoda soft-computingu, strojové učení, soubory klasifikátorů, nákladově citlivé učení

Table of Contents

LIST OF SYMBOLS AND ABBREVIATIONS

Acc	Accuracy
AdaBoost	Adaptive boosting
AI	Artificial intelligence
AMD	Advanced micro device
AUC	Area under the curve
BDT	Bagged decision tree
BN	Bayes net
BPNN	Back propagation neural network
BWM	Best-worst method
BOW	Bag of words
CART	Classification and regression trees
CCR	Corporate credit rating
CBR	China bond rating
CBDT	Clustering-based decision tree
CEO	Chief executive officer
CR	Credit rating
CRA	Credit rating agency
CRSP	Center for research in security prices
COMPUSTAT	Computer statistics
CNN	Convolutional neural networks
CSR	Corporate social responsibility
DECORATE	Diverse ensemble creation by oppositional relabeling of artificial training examples
DevOps	Development operations
DNN	Deep neural network
Doc2Vec	Document to vector
DT	Decision tree
EDA	Exploratory data analysis
ELM	Extreme learning machine
ERT	Extremely randomized trees
ESG	Environmental, social and corporate governance
EMMA	Electronic municipal market access
FMD	Financial market development
FN	False negative
FNR	False negative rate
FP	False positive
FPR	False positive rate
FS	Feature selection
FX	Foreign exchange
GB	Gradient boosting
GCC	Gulf cooperation council
GDP	Gross domestic product
GFOA	Government financial officers association
GPU	Graphics processing unit
HPC	High-performance computing
HY	High yield
IBM	International business machines corporation

IG	Investment grade
IO	Input/output
KDD	Knowledge discovery in database
k-NN	*k*-nearest neighbours
LDA	Linear discriminant analysis
LIME	Local interpretable model-agnostic explanations
LogitBoost	Logistic boosting
LSTM	Long short-term memory
LR	Logistic regression
MBSVM	Modified boosted support vector machine
MC	Misclassification cost
MCC	Matthews correlation coefficient
MCDM	Multi-criteria decision-making
MDA	Multiple discriminant analysis
MD&A	Management's discussion and analysis
MDL	Minimum description length
ML	Machine learning
MLP	Multilayer perceptron network
MSA	Multi-head self-attention
MSRB	Municipal securities rulemaking board
MTL	Multitask learning
MXNET	Mixed precision numerical computing
NB	Naïve Bayes
NFS	Network file system
NN	Neural network
NVMe	Nonvolatile memory express
NWC	Net working capital
OB	Operating balance
OR	Operating revenue
OLR	Ordinal logistic regression
OLS	Ordinary least square
PCA	Principal component analysis
POSIX	Portable operating system interface
PPP	Purchasing power parity
PSO	Particle swarm optimization
PLTR	Penalized logistic tree regression
PNN	Probabilistic neural network
P2P	Peer to peer
RBF	Radial basis function
REPTree	Reduced error pruning tree
RF	Random forest
RIPPER	Repeated incremental pruning to produce error reduction
ROC	Receiver operating characteristics
RSS	Random subspace
RUX	Rule extraction
SCB	Scheduled commercial bank
SCR	Sovereign credit ratings
SDP	Semidefinite program
SHAP	Shapley addictive explanations

SHMM	Student's *t* hidden Markov model
SMB	Server message block
SME	Small and medium-sized enterprises
SMO	Sequential minimal optimization
SMOTE	Synthetic minority oversampling technique
SPSS	Statistical package for social sciences
SSDs	Solid-state drives
SVM	Support vector machine
SVR	Support vector regression
S&P	Standard & Poor's
tf.idf	Term frequency-inverse document frequency
TE	Total expenditure
TELs	Tax and expenditure limitations
TEJ	Taiwan economic journal
TN	True negative
TNR	True negative rate
TOPSIS	Technique for order preference by similarities to ideal solution
TP	True positive
TPR	True positive rate
TR	Total revenue
US	United States
USA	United States of America
WEKA	Waikato environment for knowledge analysis
Word2Vec	Word to vector
XGBoost	Extreme gradient boosting

Introduction

Undoubtedly, the term 'regional financial performance' refers to the economic well-being of specific sub-sovereign entities, including cities, municipalities, counties, or states/provinces. This performance is commonly gauged through various financial indicators that mirror the region's economic activities and overall financial health, as highlighted in studies by Buendía-Carrillo et al. (2020), Lukac et al. (2021), Ni et al. (2023), Faridi et al. (2023), Feng et al. (2023), Ghosh et al. (2023) and others. These entities, possessing a degree of decision-making autonomy, are pivotal in managing infrastructure, delivering public services, and tax collection within their jurisdictions, thereby significantly contributing to national economic growth and development (Carmeli, 2003; Wu et al., 2021).

The increasing shift of decision-making from central to local governments underscores the importance of assessing their financial performance. As defined by Cohen et al. (2012), regional financial performance is the capacity of a region to fulfill its financial obligations and commitments to its citizens both presently and in the future. This performance serves as a cornerstone for decision-making, offering critical insights for investment opportunities. Furthermore, it lays the foundation for decision-making by offering substantial evidence on areas of investment interest (Gousario and Dharmastuti, 2015). Therefore, regional government stakeholders are keen on having indicators that forecast potential financial difficulties within the region. However, with limited exceptions, regions cannot declare bankruptcy, which complicates the evaluation of regional financial performance. Although financial distress prediction models are extensively utilized to assess corporate financial health (Liang et al., 2020; Sun et al., 2021), the development of similar models for non-profit entities has received scant attention, despite their significant role in regional and national development (Kloha et al., 2005; Wang et al., 2007; Zafra-Gomez et al., 2009; Cohen et al., 2012; Gorina et al., 2018; Antulov-Fantulin et al., 2021). The challenge in predicting financial distress for regions is exacerbated by the scarce instances of regional defaults. Credit rating agencies use a variety of financial indicators to assign overall ratings (Gaillard, 2006; Gaillard, 2009), merging a wide range of qualitative and quantitative data on various economic, financial, and political risks into composite risk ratings. The assessment of regional financial performance globally is influenced by numerous socio-economic factors, including government borrowing levels, national debt, economic growth prospects, debt interest payments, and the GDP percentage (Mohapatra et al., 2018).

Overall, despite the critical role of regional financial performance in national and regional development, the development of predictive models for such entities, which are predominantly non-profit, has received limited attention compared to corporate entities. The challenge is further compounded by the rarity of regional defaults, although the risk of financial distress for sub-sovereign entities is monitored by credit rating agencies like Moody’s and Standard and Poor's, which utilize various financial indicators to assign ratings (Ioannou et al., 2021).

Recent studies by Golbayani et al. (2020), Wang and Ku (2021), Sun et al. (2022), Wu et al. (2022), and Kumar Roy et al. (2023) on the use of ensemble learning and soft computing methods for predicting corporate credit ratings have yielded impressive results when compared to traditional statistical methods. Zadeh originally coined the term 'soft computing' in 1994, defining it as a consortium of computational techniques rooted in artificial intelligence (AI), which includes fuzzy systems, evolutionary computing, neural computing, and probabilistic methods (Zadeh, 1994). In contrast, hard computing prioritizes precision, certainty, and rigor, leading to deterministic outcomes (Chakraborty et al., 2017). Within the realm of financial distress prediction, Kumar and Ravi (2007) have classified soft computing methods into three main categories: (1) classification methods enhanced by intelligent feature selection, (2) integrated hybrid systems, such as fuzzy neural networks or evolutionary neural networks, and (3) ensemble classification methods. Similarly, Sun et al. (2004) organized these methods into: (1) classification methods paired with feature selection, (2) the use of one method to refine another classification method, and (3) the creation of a new classification method through the integration of two or more existing methods.

Notably, the predictions from multiple classifiers can be combined to produce a more robust prediction model. Ensemble methods, in particular, combine the forecasts of several base machine learning algorithms, thereby enhancing the accuracy and robustness beyond what single algorithms could achieve in predicting corporate financial distress (Zhang et al., 2022). Both theoretical and empirical studies have demonstrated that an effective ensemble is characterized by each classifier within the ensemble being precise and committing errors on distinct segments of the input space (Optiz and Maclin, 1999). This efficacy underpins the popularity of ensemble classifiers in credit rating predictions (Zhu et al., 2017; Toseafa, 2018).

Boosting and bagging stand out as two fundamental approaches in ensemble learning. Boosting methodically employs a learning algorithm to iteratively produce a series of simple, modest-quality classifiers, with each addition leading to a re-weighting of instances in the training set. This re-weighting ensures that subsequent classifiers pay more attention to the more challenging examples, and each classifier is endowed with a specific voting strength. Conversely, bagging, introduced by Breiman (1966), serves as a technique to diminish variance, making it particularly suitable for algorithms that are inherently unstable and prone to high variance.

The applications of soft computing have introduced two primary advantages. First, they offer solutions to nonlinear problems where mathematical models fall short, incorporating human-like knowledge—including cognition, recognition, understanding, and learning—into computing. This has paved the way for the development of intelligent systems, such as autonomous self-tuning systems and automated design systems (Omolaye et al., 2017). Another significant benefit is the complementary nature of soft computing techniques, which, when combined, can address complex issues beyond the scope of traditional mathematical approaches (Ammar et al., 2001). Additionally, features like intelligent control, nonlinear programming, optimization, and decision support have increased the popularity of soft computing across various applications (Ibrahim, 2016).

Beyond the ensemble classification methods previously discussed, several researchers have explored integrated hybrid approaches, such as neuro-fuzzy systems (Khashman, 2009), the combination of rough sets and neural networks (Kim and Park, 2016), evolutionary fuzzy systems (Hajek, 2018), evolutionary neural networks (Pan, 2012), the integration of rough sets and support vector machines (Yeh et al., 2010), fuzzy systems with support vector machines (Wang et al., 2005), and the fusion of case-based reasoning with support vector machines (Sun et al., 2018). These integrated methods leverage the strengths of various soft computing techniques while mitigating their weaknesses. For instance, the combination of neural networks and fuzzy systems enhances both the accuracy and interpretability of the resulting models.

Inspired by the findings discussed, this thesis is dedicated to the development of an innovative model for predicting regional financial performance, utilizing a blend of soft computing techniques. This approach marks a departure from previous studies that primarily relied on single statistical or machine learning methods. By integrating feature selection, data balancing and ensemble learning, this work aims to harness their collective strengths, offering a robust solution for forecasting the complex, uncertain, and imbalanced nature of regional financial performance.

1 State-of-the-art in Forecasting Regional Financial Performance

1.1 Importance of Regional Financial Performance

Regional financial performance reflects the economic and financial health of specific geographic areas, such as cities, states, provinces, or countries. It encompasses the analysis and evaluation of various economic indicators, financial metrics, and trends within a particular region. Understanding regional financial performance is crucial for businesses, investors, policymakers, and other stakeholders, offering valuable insights into the economic conditions, opportunities, and challenges of the area (Leiser and Mills, 2019; Maher et al., 2020; Buendía-Carrillo et al., 2020; Lukac et al., 2021).

Financial distress represents a significant concern within the public finance and economic systems of regional and local governments, particularly in the aftermath of the 2008 financial crisis. Notable examples include Detroit's 2013 bankruptcy filing, the largest in U.S. local government history, and significant financial distress cases in Catalonia, Portugal, and Italy (Gregori and Marattin, 2019; Antulov-Fantulin et al., 2021). Research by Antulov-Fantulin et al. (2021) employing machine learning and statistical models has demonstrated the feasibility of accurately predicting local government financial distress with a high true positive rate and a low false positive rate.

Moreover, the assessment of credit ratings for both sub-sovereign (regional) and sovereign (national) entities is vital, affecting not only the governments themselves but also other issuers within the region or country, including banks, companies, and public sector entities. Sovereign ratings essentially set the borrowing ceiling for these entities (Gennaioli et al., 2018). With the increasing globalization of financial markets, the demand for sovereign and sub-sovereign ratings has surged, influencing fund management strategies based on international market dynamics (Liu and Tan, 2009). Consequently, shifts in these ratings can have significant repercussions on the re-evaluation of international markets.

Bhatia (2002) elucidates that sovereign credit ratings, issued by credit rating agencies, evaluate the capability and willingness of governments to meet their commercial debt obligations promptly and in full. These ratings also furnish investors with a clearer insight into the risk level associated with investments in a specific country. Sovereign ratings play a pivotal role as they influence operations in the domestic market. The sovereign credit rating serves as a key measure for gauging the credit risk associated with numerous assets within a country. Hence, maintaining objectivity and eschewing subjective judgments in assigning sovereign credit ratings is imperative due to their significant influence on corporate and financial ratings. Borensztein et al. (2007) emphasize the importance of a sovereign credit rating, noting its substantial effect on the local market, encompassing businesses and their stock prices. In a shift from past practices, credit rating agencies have adopted a policy of not rating firms higher than the sovereign rating, which traditionally set a benchmark for the debt issued by local companies. Nonetheless, sovereign ratings primarily assess the credit risk of national governments and do not directly reflect the default risk of other issuers. Beck et al. (2017) points out that credit ratings are commonly employed to represent the financial health and arrangements of sub-sovereign and municipal entities, often relying on the sovereign's credit rating. This dependency on rating agencies renders the national financial system susceptible to vulnerabilities. A downgrade of a sovereign credit rating to below investment grade can have catastrophic effects, potentially triggering forced liquidations and significant price declines, known as cliff effects (Eijffinger 2012; Zhang and Chi, 2018; Olowookere and Adewale, 2020).

Sovereign and sub-sovereign credit ratings account for a small fraction of the rating industry in terms of the number of entities assessed. Yet, their influence on financial markets is profound. According to Amstad and Packer (2015), rating a region is comparable to capturing a snapshot of its financial, economic, and political landscape at a specific moment. Sovereign and sub-sovereign credit ratings offer an alpha-numeric depiction of the likelihood that the issuer will meet its obligations fully and punctually. It is crucial to recognize that changes in sovereign credit ratings can impact the interest rates of assets in other countries due to economic and financial linkages, thereby exerting a significant effect both domestically and internationally (Alsakka and Gwilym, 2013).

Research efforts by authors such as Trevino and Thomas (2002) and Aktan et al. (2019) have explored the modeling of sovereign credit ratings using ordered logistic regression and simple linear regression methods. In the realm of sub-sovereign (regional and municipal) credit ratings, Gaillard (2009) employed ordered logistic regression to forecast sub-sovereign credit ratings for the subsequent year, identifying default history, the ratio of net direct debt to operating revenue, and GDP per capita as key determinants. Lara-Rubio et al. (2017) pinpointed population, debt composition, and per capita income as crucial factors influencing local government credit ratings. Navarro-Galera et al. (2017) suggested that political factors might also serve as significant predictors of local government credit ratings. Despite these efforts to elucidate the determinants of regional financial health, previous studies have faced limitations due to scarce data, a focus on single countries, and the reliance on linear statistical models that fall short of fully capturing and forecasting the intricate economic and financial dynamics impacting overall financial health in the future.

1.2 Corporate Credit Rating Forecasting

The credit rating process involves a subjective evaluation that incorporates both quantitative and qualitative factors, encompassing the attributes of entities, industry dynamics, and market conditions. It initiates with a request from the issuer or the entity in question to a rating agency for assessing new debt or other financial instruments. This process is structured according to established academic conventions, including standard sections and the customary formatting of author and institutional affiliations. The issuer furnishes the rating agency with essential documentation, such as financial statements, preliminary official statements, and prospectuses for the debt issue, alongside other pertinent non-financial information. Subsequent to this, dialogues between the entity's management and the rating agency occur, culminating in the preparation of a rating report by credit analysts who scrutinize the entity under review. The credit analyst then proposes a credit rating to a rating committee, which bears the responsibility of determining the final rating to be assigned (Hajek and Olej, 2011).

In the realm of corporate credit rating modeling, the objective is to devise models capable of forecasting the credit ratings of companies by analyzing a multitude of factors, including financial data, industry performance, and broader economic indicators. This modeling process typically employs statistical methods and econometric analyses to generate precise and dependable predictions of credit ratings. Such models are instrumental for investors and lenders in gauging the risk entailed in extending credit to or investing in specific companies. Corporate credit ratings are of paramount importance in the financial domain, offering crucial insights into a company's financial stability and aiding stakeholders in making well-informed decisions. These ratings reflect the risk level associated with financial engagements with a company, facilitating investors in evaluating the probability of timely interest and principal repayments. Furthermore, lenders leverage these ratings to dictate the terms of loans, including interest rates and collateral requirements (Hirk et al., 2022).

Therefore, credit rating modeling is concerned with the creation of statistical and machine learning models to evaluate the creditworthiness of individuals, companies, or other entities. The aim is to forecast the potential for default or credit risk based on a variety of financial and non-financial parameters. Credit rating models are vital tools in the financial industry, aiding lenders, and credit agencies in making well-informed credit or loan decisions (Hajek, P., 2011).

Several studies on credit rating modeling, such as those by Golbayani et al. (2020, 2021) and Camanho et al. (2022), have conducted comprehensive investigations and comparative analyses of findings from various literature sources that employed machine learning approaches to forecast credit ratings. Four machine learning algorithms—bagged decision trees (bagging), random forest, support vector machine (SVM), and multilayer perceptron (MLP)—were analyzed, revealing outcomes that have been proven valuable in prior research on the same datasets. The experimental findings suggest that decision tree-based models outperformed other models when applied to the selected datasets. Additionally, Ubarhande et al. (2021) undertook an extensive review of the literature concerning various aspects of credit rating, which refers to the assessment of an individual's or entity's creditworthiness. Credit-rating agencies (CRAs) are entities that evaluate and assign credit ratings based on borrowers' debt repayment capabilities. A credit-rating model is a tool used by CRAs to determine these ratings. Their analysis of 153 publications aimed to identify gaps in the credit-rating field and propose solutions to address these deficiencies. A significant portion of the research, particularly that emerging from the financial crisis between 2008 and 2016, found that 48% of the studies focused on developing new credit-rating mechanisms without thoroughly evaluating the existing frameworks.

Furthermore, Zhang and Chi (2018) investigated loan clients with diverse credit ratings to determine if lower credit ratings correlated with higher loss rates and if the distribution of consumers followed a bell-shaped curve. They employed a multi-objective programming approach to develop their credit rating model, with one objective function aimed at minimizing the discrepancy between the percentage of obligor numbers and the ideal fraction of clients, assuming a typical normal distribution. The second objective function sought to minimize the overall discrepancy between the loss rates of adjacent credit ratings. Their research, which analyzed data from 6,155 firms sourced from a Chinese bank and Prosper peer to peer (P2P) loan data, demonstrated that their proposed method could effectively balance both criteria and avoid excessive concentration of obligors in certain grades.

Moreover, Camanho et al. (2022) explored the effect of competition among CRAs on the balance between maintaining a good reputation (which leads to future revenue) and engaging in rating inflation (which generates immediate revenue). Their findings suggest that, in a duopoly compared to a monopoly, rating agencies are more inclined to inflate ratings artificially. They concluded that reducing entry barriers, thereby allowing CRAs with lesser reputations to enter the market, could potentially lead to increased rating inflation and a decline in overall welfare.

Wu et al. (2022) explored the influence of supply chain information on forecasting enterprises' credit ratings. Their study, leveraging firm-level supplier-customer connections and corporate credit rating data, utilized a machine learning framework based on gradient boosted decision trees to assess the impact of supply chain characteristics on credit rating prediction accuracy. The research aimed to discern specific supply chain links that significantly enhance predictability due to their rich informational content. Similarly, Yu et al. (2022) assessed the performance of various machine learning models in predicting credit ratings of environmentally conscious companies, analyzing a sample of 355 Eurozone enterprises graded on their climate change scores by semidefinite program (SDP) from 2010 to 2019. The findings highlighted that classification and regression trees were most accurate in predicting credit ratings.

Roy et al. (2023) proposed a comprehensive fuzzy credit rating model to address deficiencies in existing models. They employed the fuzzy best-worst method (fuzzy-BWM) for weighting critical factors affecting creditworthiness and the fuzzy approach for order of preference by similarity to ideal solution (fuzzy-TOPSIS)-Sort-C for borrower assessment, integrating TOPSIS-Sorting with fuzzy theories to mitigate human ambiguity in decision-making. Doumpos and Figueira (2019) also delved into the development of internal credit rating models based on expert judgment within a multi-criteria classification framework. Their analysis assessed the models' internal characteristics and their alignment with external benchmarks provided by rating agencies.

Li et al. (2019) used an event-study approach to examine the immediate impact of credit rating announcements on financial markets, finding that market responses to credit rating releases are heterogeneous and can vary between positive and negative reactions. Their research aimed to experimentally explore the linear and non-linear effects of credit ratings on financial market development in Europe, employing the autoregressive distributed lag model to incorporate asymmetries in credit rating changes. Sajjad and Zakaria (2018) studied the role of credit ratings in capital structure decisions among non-financial Asian listed firms, using various econometric approaches to uncover a non-linear, inverted U-shaped relationship between credit rating scales and leverage ratio. Hu et al. (2019) investigated the response of incumbent issuer-paid CRAs in China to the establishment of China Bond Rating (CBR), an independent agency. The entry of CBR into the market led to a reduction in rating inflation and increased the informativeness of rating changes among existing CRAs.

Teixeira et al. (2018) analyzed the determinants of sovereign credit ratings using panel data from 86 countries from 1993 to 2013. The study examined regional variations in average credit ratings during crisis and non-crisis periods, and the impact of significant global events on these ratings, found that sovereign credit ratings are influenced by a variety of macroeconomic and qualitative factors, with noticeable differences across geographical regions.

Reusens and Croux (2017) carried out a comparative analysis to understand the significance of various factors affecting sovereign credit ratings over time, utilizing data from 90 countries spanning from 2002 to 2015. They employed the composite marginal likelihood approach to estimate a multi-year ordered probit model for each of the three leading credit rating agencies. The findings revealed that, following the onset of the European debt crisis in 2009, the importance of fiscal balance, economic growth, and external debt markedly increased. Consequently, eurozone membership's impact shifted from positive to negative. Notably, GDP growth became significantly more important for countries with high levels of debt, while government debt grew in importance for countries experiencing low GDP growth rates. This study provides solid evidence that credit rating agencies adjusted their assessment criteria for sovereign credit ratings in response to the European financial crisis.

Similarly, Takawira and Mwamba (2020) investigated the determinants of sovereign credit ratings, analyzing macroeconomic indicators data, including Sovereign Credit Ratings (SCRs), from 1999 to 2020. Their analysis aimed to identify the criteria used by CRAs to evaluate a country's ability to repay its debt and to predict future ratings. CRAs assess a wide array of factors, encompassing political, infrastructural, financial, and economic aspects, among others, of a country. This information is synthesized into a rating system, where a higher rating denotes greater creditworthiness, and a lower rating indicates a higher likelihood of default. The study focused on the three major credit rating agencies: Fitch, Moody's, and Standard & Poor's, highlighting the diverse macroeconomic data and methodologies CRAs use to evaluate and assign credit ratings to sovereign entities.

Lee et al. (2021) explored the changes in credit ratings made by two different types of credit rating agencies: an investor-paid CRA, Egan-Jones Ratings Company, and an issuer-paid CRA, Moody's Investors Service. The study aimed to assess the influence of conflicts of interest and reputation on these rating changes. A novel distribution dynamics method was employed to calculate the probability distribution and the probabilities of both downgrades and upgrades in credit ratings provided by these agencies, which operate under distinct compensation models. The analysis drew upon data from 750 U.S. issuers spanning from 2011 to 2018, a period following the Dodd-Frank Act's implementation. The findings suggest that investor-paid ratings are more likely to be downgraded than issuer-paid ratings, particularly in lower rating categories. This indicates that investor-funded agencies may adopt a stricter stance toward issuers at higher risk of default to protect their reputation. Similarly, Choi et al. (2020) investigated the potential of qualitative data from companies' annual reports in predicting corporate credit ratings. The study applied three document vectorization techniques—Bag-Of-Words (BOW), Word to Vector (Word2Vec), and Document to Vector (Doc2Vec)—to transform unstructured textual data into numerical vectors, making it suitable for analysis by Machine Learning (ML) algorithms. Specifically, the research utilized the Management's Discussion and Analysis (MD&A) section from 10-K financial reports, alongside financial metrics and business credit rating data. The results from a series of multi-class classification tests revealed that predictive models incorporating both financial metrics and vectors derived from MD&A content outperformed benchmark models relying solely on traditional financial variables.

Aktan et al. (2019) investigated the effects of actual changes in credit ratings on capital structure decisions. The study utilized three models to examine the relationship between credit ratings and capital structure choices, focusing on hypotheses related to wide rating bands, notch ratings, and the distinction between investment-grade and speculative-grade ratings. Using multiple linear regression models, the research assessed these dynamics. The results suggest that firms tend to issue less net debt relative to equity following a change in their overall credit rating level, such as a downgrade from A- to BBB+. The findings also reveal that companies appear to be relatively indifferent to changes within notch ratings, provided they remain within the same broad credit rating category. Additionally, Zhang et al. (2019) explored credit rating management in the context of energy trading across microgrids, considering factors like transmission losses and wheeling costs. The study aimed to limit the opportunistic behavior of retailers to encourage active engagement from both consumers and retailers in energy trading. By establishing a game model with retailers as leaders and consumers as followers, the research introduced a scorecard model based on logistic regression to evaluate the credit ratings of retailers. A 'trust degree' concept was implemented as a penalty for retailers, correlating their credit ratings with potential profit reductions. This approach allowed the researchers to theoretically demonstrate the existence of a unique equilibrium for the dynamic game model. Furthermore, a best response strategy was proposed to achieve equilibrium between consumers and retailers iteratively.

Iyer et al. (2022) introduced an innovative blockchain-based system designed to support a bond-pays model in the credit rating industry, addressing the conflict of interest inherent in the traditional issuer-pays model, which has led to rating shopping and inflation. The study reviews the current practices in the credit rating sector that have contributed to numerous instances of rating failures. It proposes a new set of procedures leveraging blockchain technology to enable the adoption of a more impartial bond-pays model. To demonstrate the technological feasibility of implementing a segment of the proposed model within a blockchain framework, a proof-of-concept system named 'Rating Chain' was developed.

In a separate study, Li et al. (2020) applied machine learning techniques, a subset of artificial intelligence, to predict the credit ratings of banks in the Gulf Cooperation Council (GCC) region. The analysis was based on a dataset incorporating both macroeconomic and bank-specific variables, collected quarterly from 2010 to 2018. An out-of-sample prediction was also conducted for a subsequent three-year period. The findings revealed that the random forest algorithm achieved the highest level of precision, as evidenced by its F1 score, specificity, and accuracy scores. The predictive models maintained robust performance across all credit quality categories, from the highest credit quality to default mode.

A summary of previous studies on corporate credit rating modelling is presented in Table 1, showing the methods and datasets used and performance achieved.

Table 1:Summary of previous studies on corporate credit rating modelling

Study	Classification method	Dataset	Performance
Golbayani, Florescu & Chatterjee (2020)	bagged decision tree (BDT) random forest (RF) multilayer perceptron (MLP) SVM	The input data set covers these corporate historical financial variables from 1990–2018 for financial sector and from 2009 to 2018 for energy and healthcare sectors. These variables are taken from both Bloomberg and Compustat	Acc for Financial sector: BDT = 84.21% RF = 82.83% MLP = 73.95% One vs. One-SVM = 42.12% One vs. All-SVM = 40.14% Energy sector: BDT = 82.11% RF = 84.45% MLP = 78.19% One vs. One-SVM = 75.31% One vs. All-SVM = 59.17% Health sector: BDT = 83.90% RF = 82.97% MLP = 76.63% One vs. One-SVM = 71.29% One vs. All-SVM = 61.89%
Munoz-Izquierdo et al. (2022)	decision tree (C4.5) PART algorithm rough set theory logit model	131 listed firms in the Spanish continuous trading market in 2017 with available financial data	The PART algorithm achieves the highest classification accuracy of 74.14%, followed by the C4.5 decision tree (73.28%), the rough set algorithm (72.70%) and the logit model (69.90%).
Galil, Ami and Rosenboim (2023)	Classification and regression trees (CART) and support vector regression (SVR).	COMPUSTAT database from 2005 to 2016, with an S&P issuer rating (non-default) on the financial year's last day	CART model was best with three variables, achieving Acc of 67.6%
Petropoulos et al. (2018)	Boruta algorithm (random forest model), Extreme Gradient Boosting (XGBoost) and deep neural networks (MXNET), logit model and linear discriminant analysis (LDA)	Loan level information on corporate and SME loans of the Greek banking system, from the supervisory database of the Central Bank of Greece. The dataset covers the 2005-2015 period; a 10 years' period with semi-annual information (i.e., semi-annual snapshots).	AUC for Logit = 66% LDA = 65% XGBoost = 78% MXNET = 72%
Wu, Hu and Huang (2014)	Support vector machines (SVM) Neural network (NN) Decision tree (DT) Bayesian networks (BN)	The research samples were collected from the TEJ data set. The TEJ was founded in April of 1990 in Taiwan.	DT, Bagging-DT, and Vote-(DT + BN + NN + SVM) achieved 77.38, 82.96, and 77.81 percent improvements in accuracy, respectively. Moreover, the prediction accuracies of the DT, Bagging-DT, and Vote-(DT + BN + NN + SVM) reached 86.90, 92.99, and 88.64 percent, respectively, in 1-away evaluation

Study	Classification method	Dataset	Performance
Viswanathan et al. (2020)	Modified boosted SVM (MBSVMs), SVM, neural networks, discriminant analysis (LDA) and *k*-nearest neighbors (*k*-NN) classification	23 Scheduled Commercial Banks (SCBs) have been chosen (in India)	The MBSVM exhibited the highest geometric mean (GM) values in all the three FY data consistently with an average value of 94.76%. Next to MBSVM, the SVM with RBF kernel performed well and secured the second spot with an average GM of 89.18%.
De Oliveira and Montes (2023)	*k*-nearest neighbors, gradient boosted random trees and MLP methods	Monthly dataset from January 1996 to November 201	MLP technique is the most reliable one. Its predictive accuracy is relatively high if compared to the other two methods.
De Paulo et al. (2019)	Random forest model, logit and ordinary least squares method (OLS) regression	Brazilian credit union. Behavioral and demographic variables of loans were observed in a 24-month period (from January 2015 to December 2016), totaling 2,012 observations.	AUC for logistic regression = 0.94 random forest = 0.96 Kolmogorov–Smirnov for logistic regression = 0.69 random forest = 0.90 Gini coefficient for logistic regression = 0.88 random forest = 0.92
Shengdong, Yunjie and Jijan (2023)	Stacking algorithm	Real data from an entrepreneurial borrowing	The performance of the Stacking-based model migration learning is further improved compared to the benchmark model without migration learning techniques, with the model AUC value rising to 0.8.

A corporate credit rating represents an evaluation conducted by an independent organization, assessing the likelihood of a firm meeting its financial obligations on time. These ratings, carried out by credit rating agencies, gauge a company's ability to repay its debts, offering an insight into its financial health (Tuovila, 2021). The process involves categorizing businesses based on their creditworthiness, where each company is assigned a grade according to a specific rating scale, reflecting its credit standing (Hajek and Michalak, 2013). Corporate credit ratings encompass a wide range of information, including financial performance, operational characteristics, and the broader business and economic environment in which the company operates (Doumpos and Figueira, 2018). Regulatory bodies can leverage these ratings to ascertain regulatory capital requirements with greater precision (Gama and Geraldes, 2012). Essentially, a corporate credit rating is a subjective judgment by a team of experts who assess the issuer's overall financial strength and ability to honor its financial commitments. It serves as a formal and unbiased indicator of a company's debt repayment capacity. These ratings are widely utilized by various stakeholders, such as investors and the companies themselves, to guide their decision-making processes (Goldmann, 2022).

Extensive research has been undertaken in the domain of corporate credit ratings. Notably, Kaur et al. (2023) performed a bibliometric analysis on the determinants of corporate credit ratings, reviewing 135 papers published from January 2001 to June 2021. This analysis aimed to pinpoint significant themes, topics, and the conceptual, intellectual, and social structure underpinning the body of knowledge through various bibliometric techniques. The findings underscore the interdisciplinary nature of the research and its theoretical foundations in bankruptcy studies. Furthermore, the study highlighted emerging research areas such as corporate social responsibility (CSR), environmental, social, and corporate governance (ESG) criteria, machine learning, management expertise, and sustainability within this field. Similarly, Chi et al. (2020) explored the differences in CSR disclosure practices between private and public companies. Their findings suggest that private firms are less likely to issue CSR reports compared to their public counterparts. Employing a bivariate probit model to accommodate partial observability, the study concluded that supply-side forces predominantly drive these differences, rather than demand-side influences. From the perspective of creditors, public companies benefit from enhanced credit ratings and lower borrowing costs due to their CSR disclosures. However, private firms do not enjoy similar benefits from CSR activities. The study also found that robust corporate governance and CSR assurance can alleviate debt holders' concerns regarding private companies' CSR engagements. Lin et al. (2020) explored the influence of corporate governance and CSR on the credit ratings of Taiwanese companies. Utilizing ordered logit regressions with two-stage least-squares estimates, the study examined the causal relationship between these factors. The findings reveal that CSR performance plays both a moderating and a partially mediating role in the relationship between corporate governance and credit ratings, suggesting that effective governance and active CSR engagement can positively impact a firm's credit rating.

Petropoulos et al. (2016) introduced an innovative approach for corporate credit rating using Student's *t* Hidden Markov Model (SHMM), which are adept at modeling heavy-tailed time-series data. This method employs a selected set of financial ratios for credit scoring, subsequently modeled with SHMM. The effectiveness of this approach was evaluated using a dataset pertaining to Greek businesses and SMEs, encompassing five years of financial data and instances of delinquent behavior. Comprehensive comparisons were made between the credit risk assessments derived from this method and those from models traditionally employed by financial institutions. The proposed approach yielded highly accurate forecasts, offering a crucial, intelligent tool for banking professionals to improve their decision-making processes. Furthermore, Gupta (2023) delved into the determinants of credit ratings assigned to Indian enterprises, focusing on both long-term and short-term perspectives. Utilizing panel data and cross-sectional analysis, the study examined the impact of various factors on credit ratings. An ordered probit analysis was conducted to explore the relationship between credit ratings (the dependent variable) and six financial metrics (the independent variables). The analysis revealed a strong correlation between a company's size, profitability, and leverage, and its corporate credit ratings in both panel data and cross-sectional evaluations. Notably, company size emerged as the most significant factor influencing credit ratings, followed by leverage and profitability. The major contribution of this research lies in the development of two distinct mathematical models that demonstrate a high prediction accuracy. These models can be utilized by investors, academics, practitioners, and other stakeholders to predict the rating categories of various firms, providing a reliable insight into their creditworthiness and financial stability.

Galil et al. (2023) employed machine learning methods, specifically Classification and Regression Trees (CART) and Support Vector Regression (SVR), to predict corporate credit ratings. While SVR showed slightly higher accuracy, CART was noted for its interpretability. However, a challenge with unrestricted models is their potential to create non-monotonic relationships between credit ratings and fundamental characteristics, an undesirable outcome. To address this, the study advocated for the use of constrained CART models, ensuring interpretable and theoretically sound results. The research underscored the importance of firm size in the accurate prediction of credit ratings and proposed an optimal model incorporating size, interest coverage, and dividends as key variables.

Al-Najjar and Al-Najjar (2014) developed an NN model to predict the credit ratings of non-financial firms in Jordan, utilizing 19 financial indicators such as profitability, leverage ratios, liquidity, bankruptcy risk, and sales performance. The study applied two neural network approaches: the Kohonen network and the Back Propagation Neural Network (BPNN). The BPNN algorithm effectively distinguished between high-performing (A-rated) and bankrupt (D-rated) companies during the 2005-2007 period. Furthermore, Feng et al. (2020) evaluated the performance of traditional machine learning models in predicting corporate credit ratings and introduced a novel approach called Corporate Credit Ratings via Convolutional Neural Networks (CCR-CNN). This method capitalizes on the strengths of convolutional neural networks and extensive financial datasets, transforming each company into an 'image' from which CNNs can discern complex feature interactions that previous models might have missed. Comprehensive evaluations using a dataset of Chinese publicly-listed companies, specifically compiled for this research, demonstrated that CCR-CNN consistently outperforms existing leading methods.

Hajek et al. (2017) developed a method for extracting themes from annual reports related to corporations using latent semantic analysis. This approach integrates the extracted information with traditional financial metrics to construct a multi-class model for predicting corporate credit ratings. Informative features are identified through a correlation-based filter during the feature selection phase. The study found that Naïve Bayesian networks offer classification performance comparable to other machine learning techniques, which is statistically significant. It was also shown that "red flag" indicators identified by Naïve Bayesian networks could signal poor credit quality, particularly in non-investment grade categories for businesses. These findings hold substantial importance for investors, banks, and market regulators.

In the work of Koerniadi (2023), the research centered on the impact of changes in businesses' credit ratings on their subsequent risk-taking activities, along with the strategies employed by these firms to implement their risk-taking approaches. Using fixed-effect regression models, the study analyzed the risk behavior of companies following downgrades to the lower boundary of investment-grade rating (i.e., BBB-) and below. The findings suggest that changes in credit rating generally correlate with reduced risk-taking post-event. However, it was observed that companies downgraded to BBB- did not show an increase in risk-taking behavior.

Additionally, Zukanovic et al. (2023) critically evaluated existing credit rating methodologies employed by major agencies like Moody's, Standard & Poor's, and Fitch, and proposed an enhanced data model for predicting corporate ratings using computational intelligence. This research aims to offer new perspectives on credit rating and its predictive accuracy to academics and practitioners alike. The study focused on a select group of companies listed in the S&P 500 index, analyzing data from financial reports covering the period from 2016 to 2019, including various financial metrics. The primary focus was on designing the data model, preparing the data, and managing missing information.

Zhang et al. (2023) developed an advanced neural model to assess the creditworthiness of manufacturing companies, incorporating a Multi-head Self-attention (MSA) mechanism and a Long Short-Term Memory (LSTM) network. The model employs MSA to simulate market dynamics and assign dynamic weights to each indicator, using financial data from manufacturing enterprises. Meanwhile, LSTM extracts sequential features from comprehensive financial and operational data to understand the long-term financial health and minimize the risk of deviation. The experimental results indicate that this methodology provides more unbiased and reliable credit ratings for manufacturing companies.

Ren et al. (2015) aimed to predict business credit ratings using various data mining techniques, focusing on the use of SVM and more advanced methods such as SVM+ and SVM+MTL (Multi-Task Learning). The study showcased the effectiveness of these novel strategies in multi-classification and corporate credit rating prediction, comparing their performance with traditional SVM. The research addressed the challenges of multi-class SVM+ and SVM+MTL by creating multiple binary classifiers and demonstrated the superior accuracy of the SVM+MTL algorithm in predicting business credit ratings over traditional SVM and SVM+.

Gu et al. (2022) examined the impact of the geographical proximity between corporate headquarters and the main offices of credit rating agencies on the acquisition of corporate confidential information by the agencies. The study found a negative correlation between geographic distance and the level of corporate private information obtained by rating agencies. This relationship contributes to increased information asymmetry, lower credit ratings for institutional and corporate bonds, and a decline in the quality of rating information content.

Cao et al. (2019) explored how management's risk tolerance affects corporate credit ratings and the mechanisms through which this influence manifests. The study used the possession of a private pilot license by CEOs as a proxy for their risk tolerance and found that companies led by pilot CEOs tend to have lower credit ratings. This observation remains consistent even after controlling for variables such as company fundamentals, CEO incentives for risk-taking, and other CEO characteristics. Path analysis revealed that risk-prone CEOs contribute to lower credit ratings by reducing future business value, increasing the volatility of future firm value, and influencing the assessment of management by rating agencies. Additionally, the relationship between CEO risk tolerance and credit ratings becomes more pronounced when management plays a more dominant role in the company. In a related study, Cho et al. (2020) examined the effects of positive credit ratings on corporate decision-making processes. The research utilized credit ratings of Korean companies and developed a measure of credit rating optimism by comparing actual ratings to benchmark ratings based on US corporate ratings. The findings indicate that higher optimism in credit ratings is associated with lower debt costs and increased levels of debt financing and investment.

1.3 Sub-sovereign Credit Rating Forecasting

Cheikh et al. (2021) presented a groundbreaking approach to explore how sovereign ratings influence corporate ratings. In light of the recent easing of policies that previously prevented a private issuer from being rated higher than its government, further empirical investigation is essential to identify the key factors affecting the relationship between government and corporate ratings. The study introduces a nonlinear panel smooth transition regression model, allowing for the sovereign impact to vary under different financial conditions at the corporate level. The findings suggest that companies with stronger financial standings, as reflected by their interest and debt coverage ratios, are less dependent on their home country's credit risk. Furthermore, Ntsalaze et al. (2017) examined the effect of sovereign credit ratings on South African corporations, specifically questioning whether the credit ratings assigned to South Africa by rating agencies act as a ceiling for the ratings of companies within the country. The research employed a longitudinal panel design, utilizing fixed effects and generalized method of moments approaches. The main findings indicate that sovereign ratings indeed function as a cap for corporate ratings in South Africa and play a significant role in shaping business ratings. However, the study also noted that company-specific characteristics, particularly accounting variables, do not significantly explain the credit risk ratings assigned to companies.

The generation of sub-sovereign credit ratings is a complex and collaborative endeavor, requiring rating analysts to skillfully balance comparability with precision, and quantitative analysis with qualitative judgment. Municipal bond ratings offer investors insights into the likelihood of bond defaults, while sub-sovereign credit ratings influence social objectives by affecting the cost of public infrastructure projects (Yinger, 2010; Besedovsky, 2017; Omstedt, 2019). There has been considerable research in the area of sub-sovereign credit ratings. Among these studies, Gillette et al. (2020) explored how credit rating levels influence the disclosure decisions of municipal debt issuers. The study observed that following the recalibration of Moody's sub-sovereign rating scale in 2010, municipalities that received upgraded ratings significantly reduced their disclosure of essential ongoing financial information compared to those not affected by the recalibration. This reduction in disclosure was more pronounced among sub-sovereign bonds held by investors who, prior to the recalibration, relied heavily on such disclosures due to a decreased need for information among debt holders. However, the study also found that there was no reduction in disclosure among issuers when they were closely monitored by underwriters possessing detailed knowledge about the issuer or when issuers were subject to direct regulatory oversight through government funding.

Maher et al. (2016) analyzed the effect of the restrictiveness of tax and expenditure limitations (TELs) on the credit ratings of 566 U.S. municipalities from 2007 to 2010. Credit ratings from Moody's and municipal fiscal data from the Government Finance Officers Association's (GFOA) Certificate of Achievement for Excellence in Financial Reporting program were utilized. The study found that more stringent TELs imposed on municipalities modestly negatively impact their credit ratings, potentially leading to higher interest costs for these municipalities.

Herrmann (2020) explored the relationship between cities' use of data for management purposes and their credit ratings, highlighting the significant implications of credit ratings on government service provision. Employing linear regression on a unique dataset comprising city bond ratings, budgetary data, and an independent assessment of data-driven management practices, the research demonstrated that municipalities with higher credit ratings are more likely to embrace data-driven management, even after adjusting for fiscal and demographic factors identified in previous studies.

Basu et al. (2022) examined how the dissemination of information affects transaction costs in the municipal bond market, leveraging a regulatory change that made sub-sovereign credit rating data from two of the three major agencies available on the Electronic Municipal Market Access (EMMA) database. Utilizing a difference-in-differences approach, the study assessed the impact on bond trading post-regulation change, focusing on whether rating information was provided on EMMA. The findings suggest that making credit ratings widely available primarily reduced transaction costs for individual investors, particularly for bond purchases and in cases where issuer information was previously limited.

Furthermore, Sharma et al. (2023) investigated the impact of competition among CRAs on the credit rating industry's challenges. Using quantitative and regression analyses, the study assessed how competition influences a company's credit rating, drawing on financial and credit rating data from Indian companies. The research, which examined dual ratings, found that CRAs may inflate a company's credit rating due to competitive pressures, and that 'rating shopping' behavior is evident in the industry, driven by CRAs' competition for new clients. Nwogugu (2021) delved into the constitutionality of private-sector CRAs and their ratings, especially those evaluating corporate, municipal, and government financial instruments. He explored the constitutional aspects and flaws of government bailouts/bail-ins, linking them to issues in Constitutional Political Economy. Additionally, Nwogugu critiqued the unconstitutionality of various initiatives, including "Obamacare," European Union bailouts, the US auto industry bailout, and the Nigerian banking and power sectors' bailouts. While the chapter primarily addresses US law, the fundamental legal principles discussed are broadly applicable in most common law jurisdictions.

In their research, Huang et al. (2021) examined how the use of multiple credit ratings by companies changed following the Dodd-Frank Act's implementation. The study found a decreased tendency among companies to obtain a third rating, typically from Fitch, for bonds near the high yield (HY) and investment grade (IG) boundary, aiming to strengthen their new corporate bond offerings. After Dodd-Frank, the importance of third-party ratings declined, as evidenced by their reduced impact on credit spreads for firms with ratings from both S&P and Moody's that straddle the HY-IG divide. This study provides new insights into how Dodd-Frank has limited corporate borrowers' strategic use of varied credit ratings around critical thresholds.

Additionally, Hájek (2011) explored the application of NNs in accurately modeling municipal credit ratings, a challenging real-world problem. The paper begins by reviewing existing credit rating modeling methods and previous research on municipal credit rating modeling. The goal was to classify municipalities in the State of Connecticut into different rating categories. The modeling process involved data preprocessing, selecting input variables, and constructing various neural network architectures for classification. Genetic algorithms were utilized to select input variables. The results demonstrated that bond issuers' rating categories could be precisely determined using a limited set of input features.

In a study by Abakah (2020), the aim was to investigate the effect of religion-induced risk aversion on the municipal bond market from 1990 to 2017. The research indicates that U.S. counties with a higher proportion of Catholics relative to Protestants typically exhibit local government bonds with lower credit risk ratings, smaller yield spreads, and a decreased likelihood of credit enhancement. These findings remained robust upon further examination. The study also explored the impact of political party affiliations within the issuer's county and state term limits, consistently revealing significant effects. Additionally, Chun et al. (2019) developed a model to gauge the intensity of municipal yields, factoring in credit default swap premiums of insurers and data from both insured and uninsured municipal bond transactions. The model was individually estimated for 61 municipal issuers, capitalizing on the marked decline in bond insurers' credit quality from July 2007 to June 2008. Subsequent decomposition of the municipal yield spread based on the model's parameters revealed the considerable role of the liquidity component and the interaction between liquidity and default risks. This finding is consistent with the approach taken by Chen et al. (2018) in their work on assessing the liquidity and default risks of corporate bonds throughout the economic cycle.

In Liu's (2012) study, the ability of bond insurers to accurately assess the credit risk of covered bonds was scrutinized, with significant implications for the future of municipal bond insurance. The research involved analyzing a set of insured municipal bonds to determine if the premiums paid for bond insurance could predict future credit rating changes, indicative of bond credit risk. The findings suggest that while municipal bond insurance premiums, considering bond credit ratings and other pertinent factors, can explain credit rating downgrades, they do not predict upgrades.

Lorenzo et al. (2022) examined how diverse fiscal regulations across US states affect municipal bond returns. Utilizing a dynamic equilibrium model, they assessed the interplay between fiscal policies and municipal credit risk. State governments determine their optimal debt levels and default policies based on a fiscal rule that considers the government's debt and the state's economic output. A nationally recognized investor, evaluating the risk associated with municipal claims, influences the valuation of these claims. By modifying fiscal rules, the study estimated the impact of fiscal institutions on tax policy, finding that reduced fiscal stringency correlates with higher expected returns and debt levels for municipal bonds across states. This relationship is largely due to a credit risk premium, which significantly increases during economic downturns, as further supported by Masungini et al. (2023).

Table 2 summarizes existing research efforts on sovereign and sub-sovereign credit rating modelling. Similarly, as for corporate credit rating overview, methods and data, together with classification performance, are presented to provide an overview of previous research.

Table 2: Summary of studies on sovereign and sub-sovereign credit rating modelling

Study	Classification Method	Dataset	Performance
Hajek (2011)	Multiple discriminant analysis (MDA) and logistic regression (LR), genetic algorithms	Dataset from US municipalities and the other one for non-US municipalities	In this study the results obtained for four-class problem using statistical methods (78.6% for MDA, 74.4% for LR) are slightly better than previous results, while especially probabilistic NNs (98.8%) and SVMs (96.0%) achieved significantly better classification quality. For the nine-class municipal credit rating, the classification accuracies on testing data obtained by NNs are also high. The best results are achieved by PNNs (96.3%).
Shan and Nilson (2018)	Logistic regression, artificial neural network, multilayer perceptron network (MLP), SVM, decision tree, probabilistic neural network (PNN) and deep learning	The dataset consists of approved loans from 2015 – 2017, which contains 1 299 083 observations with 101 variables.	Acc for Logistics = 93.73% MLP = 93.62% Decision Tree=95.6% SVM = 97% PNN = 92.9% Deep learning = 90.0%
Pamuk and Schumann (2023)	NN, XGBoost, Logistic regression (LR), and Decision tree (DT)	For this paper, we have selected the dataset from the study of Pamuk et al. (2021), which consists of 3.3 Mio. entries of annual financial statements from 2000–2012 with 74 metrics	The ML models (NN, XGBoost, LR, and DT) are combined with four sampling techniques to balance the distribution of ten credit rating classes. The results indicate that XGBoost provides the best outcome with Synthetic Minority Oversampling Technique and Edited Nearest Neighbor (SMOTE-ENN) (75–89%).
Chalak and Kim (2022)	S&P (standard & poor) rating, Fitch, and Moody's rating	Dataset were retrieved from Municipal Securities Rulemaking Board (MSRB), Refinitiv Eikon, Bloomberg, and	Acc for S&P = 93% Moody's rating = 96.1% Fitch = 96.4%

Study	Classification Method	Dataset	Performance
		S&P Capital IQ	
Weng and Huang (2021)	C4.5 algorithm, DT (decision tree), MLP (multiple-layer perceptron), NB (naive Bayes classifiers), RF (random forest), and SVM (support vector machine). clustering-based decision tree (CBDT)	A cost matrix is provided on the Statlog (German credit data) dataset website	Acc for CBDT = 89% DT = 84% MPL = 73% NB = 75% RF = 78% SVM = 83%
Suleri (2023)	Logistics regression (LR) Penalised logistic tree regression (PLTR), RuleFit and RUle eXtraction (RUX), random forest (RF), Extremely randomized trees (ERT), Adaptive boosting (AdaBoost) and eXtreme Gradient Boosting (XGBoost)	The data that we use consists of aggregated United States (US) state-level data with LendingClub's loan book covering the period from 2008 to 2019	AUC for RF-RUX = 74.9% LR = 75.7% RF-RuleFit = 80.6% Ada-RleFit = 81.1% PLTR = 81.3% Ada-RUX = 81.7% AdaBoost = 87.8% XGBoost = 87.9% ERT = 84.1% RF = 87.2%
Cheng (2020)	Logistic regression (LR), classification and regression tree (CART), gradient boosting model (GB), SVM, random forest (RF), neural networks (NN), and semi-supervised learning	The dataset consists of approved loans from 2015 – 2017, which contains 1 299 083 observations with 101 variables	Acc for Extra Tree Classifier (ETC) = 87.35% SVM = 75% ANN = 89% RF= 81.27% DT = 95% XGBoost = 80.21% CatBoost = 80.47%
Maran, Funnell and Castellini (2019).	Logistic regression, MLP, SVM, decision tree, probabilistic neural network (PNN) and deep learning	The dataset consists of approved loans from 2015 – 2017, which contains 1 299 083 observations with 101 variables.	AUC for LR = 85.73% MLP = 83.62% DT = 87.6% SVM = 88% PNN = 72.9% Deep learning = 80.0%
Jia et al. (2021)	Principal components analysis (PCA), particle swarm optimization (PSO) and extreme learning machine (ELM) model	Data were collected from the Wind Database	PCA–PSO–ELM proposed in this research has the highest accuracy in terms of the prediction compared with ELM, BPNN and auto regression.
Nakashima, Mantovani and Machado Junior (2022)	Local interpretable model-agnostic explanations (LIME) algorithms	Dataset from Compustat S&P Ratings database between 1996 and 2013) and traditional financial variables from Center for Research in Security Prices (CRSP) databases	Average accuracy of SVM= 75.9% RF = 83.1% LR = 76.6% DT = 79.4%

1.4 Partial Conclusion

In summary, while previous research has sought to elucidate the factors influencing regional financial health, these studies have often been constrained by limited data availability (primarily focusing on single countries) and the challenge of high-dimensional data, also known as the curse of dimensionality. This issue was typically addressed by selecting the most crucial features, primarily due to the risk of imbalanced classes and the necessity to manage highly imbalanced multi-class data pertaining to regional entities. Toseafa and Hajek (2019) tackled this problem with a novel hybrid model that merges data oversampling with cost-sensitive ensemble classification, demonstrating that the Synthetic Minority Over-sampling Technique (SMOTE) can effectively balance multi-class data and mitigate the imbalance issue.

Furthermore, ensemble classification methods have gained popularity for their ability to minimize overfitting and variance. Capitalizing on these benefits, this dissertation employs potent combinations of soft computing techniques, including ensemble classifiers integrated with MetaCost classifiers and base classifiers (feature selection + classifier/hybrid methods/ensemble learning models). This approach is applied across various ensemble learning techniques, such as boosting, bagging, XGBoost, and random forest, to enhance predictive accuracy and robustness.

2 Aim and Objectives of the Dissertation

This thesis aims to employ a novel hybrid model based on the effective combination of different soft computing methods, including their ensembles, to forecast regional financial performance in terms of sub-sovereign credit ratings. Different models are compared for the real-world imbalanced multiclass classification task.

To achieve this aim, the following specific objectives are defined:

- Perform feature selection. Feature selection aims to reduce the dimension of the feature space, making learning algorithms operate faster and maximizing classification accuracy in building learning models. It is done by visualizing and understanding the data to reduce storage requirements and training times (Hajek and Michalak, 2013; Dash et al., 2000; Petrides and Verbeke, 2022). The two main models of feature selection are the filter and wrapper methods. Wrapper models optimize predictors as part of the selection process, whereas filter models rely on the general tendencies of the training data to select features that depend on any predictor. Although filter methods are usually computationally less expensive than wrappers, wrapper models tend to give better results. Correlation-based feature selection is a simple filter algorithm that ranks feature subsets according to the correlation-based heuristic evaluation function. In this thesis, I compare the effects of wrappers and filters for predicting regional financial performance.
- To demonstrate the effectiveness of the proposed combinations of soft computing methods (feature selection + classifier / hybrid methods / ensemble learning models), the results will be compared with those of the baseline and state-of-the-art machine learning methods. Notably, cost-sensitive ensemble learning, such as meta-cost, are used to overcome the problem of imbalanced data (as only few sub-sovereign entities are assessed by the best / worst rating classes). The following classification measures are used to evaluate the performance of the prediction models: Accuracy, area under ROC (receiver operating characteristics) curve (AUC) (to evaluate the performance on imbalanced classes), F-measure (the combination of precision and recall), and Misclassification cost (to consider different financial effects of misclassified regional units).
- The problem of imbalanced data is further addressed using data balancing methods, such as random oversampling and SMOTE.

- Using eXtreme Gradient Boosting (XGBoost) due to its superiority in many machine learning competitions. In broad terms, XGBoost has become popular due to its computational efficiency (parallel computation, cross-validation at each iteration), regularization through lasso and ridge methods, and feasibility (system optimization and hardware efficiency).
- Using complementary log-log approach which a statistical method that is preferred due to its asymmetrical distribution and allows for values beyond the binary classification. It is frequently used in situations where the event probability is very small or large. Furthermore, this model has a direct interpretation of hazard ratios (Allison, 2012; Gupta et al., 2018).
- Explainability of the best-performing prediction models is achieved using the SHAP values, showing the contributions and effects of sub-sovereign credit rating determinants.

3 Research Methodology

This section of the thesis presents the research methodology, as depicted in Figure 1. The process begins with the collection of benchmark datasets from Moody's credit rating agency. The subsequent step involves data preprocessing, which includes cleaning the data by filling in missing attributes, followed by normalization and aggregation to reduce the number of attributes for analysis. After preprocessing, feature selection is carried out using both wrapper and filter methods to identify a subset of features that will facilitate the development of robust learning models, such as neural networks and decision trees. The experimental phase involves training and testing the datasets with a split of 67% for training and 33% for testing, alongside the use of 10-fold cross-validation to ensure the reliability of the results. Various machine learning algorithms are employed to train a novel hybrid model. The study will also explore multiple ensemble classifiers, ensemble base classifiers, and cost-sensitive approaches.

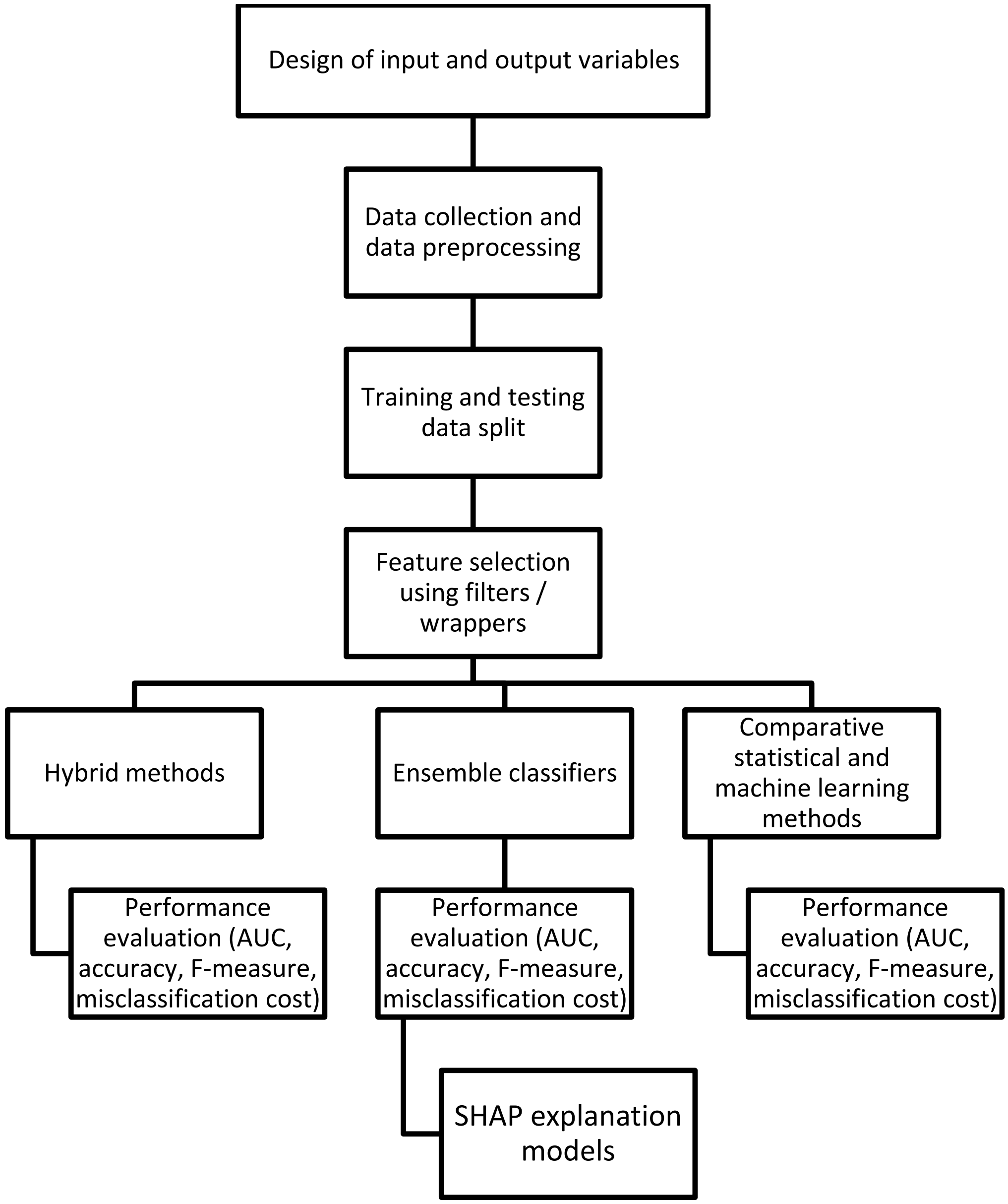


Figure 1: Research methodology

3.1 Datasets

To assess the effectiveness of credit rating prediction systems based on machine learning across different time frames, several benchmark datasets from relevant fields are utilized. Selecting datasets that showcase the broad applicability of the proposed models is crucial. As such, two categories of datasets were sourced from Moody's credit rating agency, focusing on sub-sovereign credit ratings for both developed and emerging countries across Europe, North and South America, Africa, and other regions, spanning the years 2003-2007. Additionally, data for one-year and two-year ahead forecasts for 2008 and 2009, respectively, were acquired. The input variables for credit rating evaluation were carefully chosen based on the criteria used by Moody's. The predictive model will incorporate economic, financial, and debt-related variables (as detailed in Table 3). A comprehensive time-series dataset comprising sub-sovereign entities from 257 regions during 2003-2007 was compiled based on Moody's ratings and organized into long-term rating classes ranging from Aaa, Aa1 to C.

Table 3: List of input variables for sub-sovereign credit rating prediction (dataset with 2008-2009 predictions)

Category	Input Variable	Mean	Standard Deviation
Economic	Country rating class	4.18	4.04
	Developed country (0/1)	0.54	0.50
	Government rating class	5.26	3.83
	City (0/1)	0.44	0.50
	Previous sovereign default (0/1)	0.36	0.48
	GDP per capital	21939.92	18311.96
	GDP/(national average)	105.20	44.36
	GDP in PPP	22729.18	13388.28
	Real GDP change	3.19	2.88
	Unemployment rate	6.40	3.84
Debt	Net debt per capital	188.63	3906.52
	Debt/GDP	7.06	13.76
	Debt/operating revenue	6.22	73.32
	Short-term debt/debt	16.33	22.78
	FX debt/debt	7.23	19.24
	Long-term debt/debt	51.46	41.39
	Debt maturity	7.88	3.73
Financial	Own revenue/OR	40.42	29.45
	Government transfers/OR	46.85	27.99
	Earmarked revenue/OR	23.86	25.92
	Interest/OR	2.53	2.66
	Debt service/TR	7.65	9.63
	Cash surplus/TR	-0.45	8.55
	Borrowing/TR	7.29	10.95
	TE per capita	2261.95	2819.64
	TE/GDP	12.26	7.75
	Operating balance/OR	13.49	10.89
	Gross operating balance/OR	10.99	11.14
	Self-financing ratio	0.97	1.04
	Capital spending/TE	19.72	12.91
	TR-TE(%)	0.14	2.98
	NWC/TE	4.18	22.40
Output variables	Rating class +1*	6.62	4.75
	Rating class +2	6.71	4.76

* rating classes Aaa, Aa1, … , C were transformed to numbers 1, 2, … , 21

An additional dataset, encompassing 451 regional units, was collected from Moody's for the years 2015 and 2016. These units were classified into eight rating categories, ranging from Aaa to C. The 2015 data served as the input for the predictive model, while the 2016 rating categories were used as the model's outputs. The prediction model employed the same set of variables as those detailed in Table 4, see Toseafa (2018). Figure 2 illustrates the distribution of regions across the eight rating classes, from Aaa to Ca. The histogram in the figure also highlights the significant imbalance within the dataset.

Table 4: List of input variables for sub-sovereign credit rating prediction (dataset with 2016 predictions)

Variable	Mean	Variable	Mean	Variable	Mean
Population	2848.5	Debt/OR	73.07	Fin. surplus/TR	-4.61
GDP	100,037	Net debt/TR	68.19	Cash surplus/TR	-1.83
GDP per capita	104.10	Debt before swap	9.56	Borrowing need/TR	10.28
GDP/Nat.Avg.	27,056	Debt after swap	7.82	TE per capita	4,057
GDP (PPP)	27,881	Short-term debt/Debt	15.79	TE/GDP	14.71
Real GDP	1.96	Long-term debt/Debt	50.56	OB/OR	7.39
Unempl. rate	8.527	Maturity debt	8.43	Gross OB/OR	5.04
Nat. unempl. rate	8.95	Own Revenue/OR	40.03	Net OB/OR	-1.63
Total debt	12,426	Govern. transfers/OR	40.37	Self-financing ratio	0.92
Net debt	10,895	Earmarked Revenue/OR	25.81	Capital spending/TE	17.03
Net debt per capita	3,263	Interests/OR	2.52	TR-TE	-0.04
Debt/GDP	11.86	Debt service/TR	8.17	NWC/TE	14.7

Notes: GDP – gross domestic product, OR – operating revenue, TR – total revenue, PPP – purchasing power parity, TE – total expenditures, OB – operating balance, NWC – net working capital.

The input and output variables presented in Tables 3 and 4 were derived from Moody's financial reports. Economic variables are essential in assessing the ability to generate resources for repaying sub-sovereign debt. The rating class for the country and government provides long-term foreign currency credit ratings for sovereign and government bonds, indicating the creditworthiness of these entities (Mohapatra et al., 2018).

GDP per capita serves as an indicator of relative economic performance and is included to evaluate the potential effects of redistributive programs on fiscal health. The GDP/national average reflects relative wealth levels, which can vary significantly within a country and highlight regions with concentrated economic activities. Such disparities may affect the fiscal capabilities of sub-sovereign governments. GDP in purchasing power parity (PPP) terms is used to gauge the economic size for international comparisons based on current exchange rates. However, it may not fully capture the real cost-of-living variations between countries, potentially leading to inaccuracies in representing relative living standards in certain cases. The change in real GDP, indicated by the annual percentage growth rate adjusted for inflation, is a crucial measure of economic dynamism. The unemployment rate reveals the degree of the output gap and the underutilization of the workforce, and it may also signal potential political pressure on the government to stimulate economic recovery.

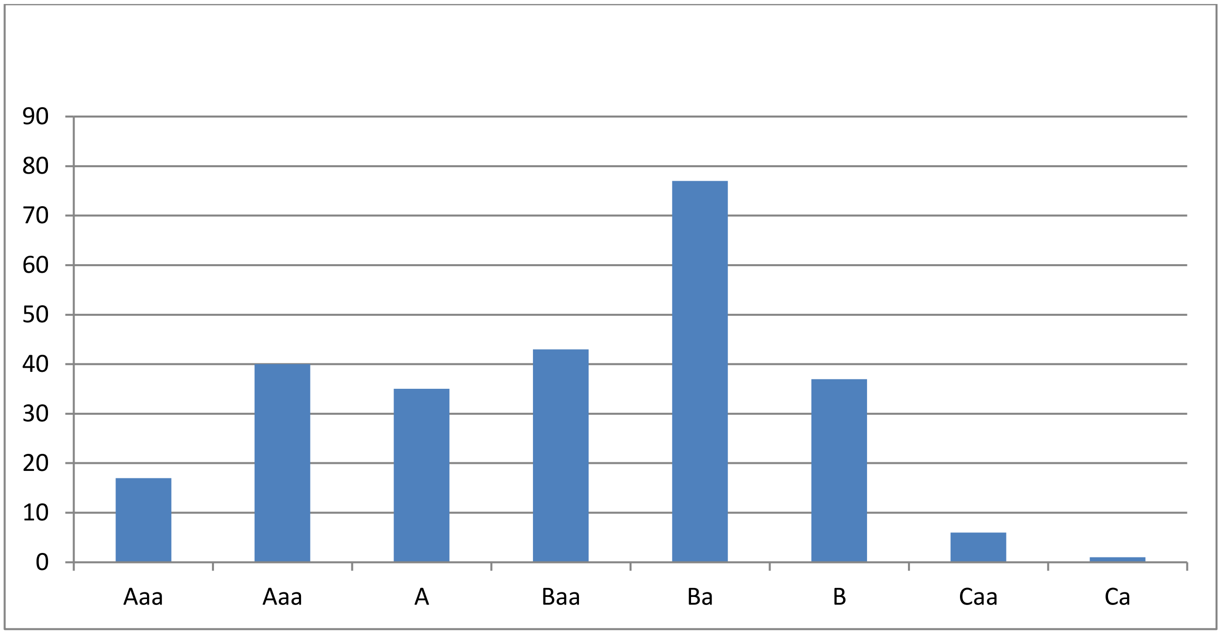


Figure 2: Histogram of sub-sovereign rating classes for 2016

Net debt per capita is a ratio that compares the sum of net direct and guaranteed debt to the population size, offering a measure of debt levels that is comparable across countries. The debt-to-GDP ratio represents the total gross government debt as a percentage of GDP, serving as a crucial indicator of government financial sustainability. Short-term debt encompasses debt instruments with a maturity of less than one year, including government securities and treasury bills, whereas long-term debt comprises instruments payable over periods longer than one year. Debt maturity delineates the timeline until the principal amount of notes, drafts, and other debt instruments are due for repayment to investors, indicating when the government's debt obligations will mature or be redeemed.

Own revenue consists of all government income from taxes, fees, and other sources under partial government control, reflecting the government's fiscal flexibility to tackle financial challenges. Governmental transfers include revenues transferred from higher government levels, intended for general fiscal support or specific spending categories. Earmarked revenue encompasses all revenue streams designated for particular purposes, excluding debt service, which can constrain a government's debt service capacity. The interest/operating revenue ratio assesses the government's capability to cover interest payments with its operating revenue. Total expenditure as a percentage of GDP gauges the impact of a government's spending on national economic output. The operating balance measures the government's inherent ability to maintain operating expenses below its operating revenues, indicating structural fiscal health. A self-financing ratio below one suggests a necessity to borrow for capital budget needs. Lastly, the NWC (Net Working Capital) ratio offers a glimpse into the entity's liquidity position, highlighting the requirement for short-term market access.[1]

[1] https://www.moodys.com/

3.2 Data Preprocessing and Feature Selection

The data preprocessing stage involved imputing missing values for some variables with their mean values, followed by normalization to the [0,1] range. Subsequently, feature selection was undertaken to streamline the number of variables for analysis. To tackle the challenge of class imbalance, the SMOTE technique was employed to generate a synthetically balanced or nearly balanced training dataset, facilitating effective classifier training (Chawla, 2002; Fernández et al., 2018). After applying SMOTE application, all classes in the training data achieved the same frequency as the most prevalent class, typically the Ba class. It is noteworthy that the testing data remained untouched during this phase, preserving its highly imbalanced nature to realistically assess classifier performance.

Given the ordinal nature of the output classes, I advocate for the use of a cost matrix, with the MetaCost classifier leveraging this matrix to enhance prediction accuracy. Consequently, MetaCost was integrated with various base classifiers, including ensemble classifiers (Toseafa and Hajek, 2019).

Feature selection, a pivotal aspect of data analysis, entails determining the most suitable subset of variables for precise prediction. The presence of redundant, irrelevant, or misleading features in data classification necessitates feature selection for effective classification problem-solving (Jirapech-Umpai & Aitken, 2005; Gheyas & Smith, 2010; Yongjun et al., 2012; Zhongyi et al., 2015). After addressing missing attributes and applying normalization and aggregation, the feature selection techniques were categorized into filters, wrappers, and embedded methods (John et al., 1994).

This analysis utilized wrapper models, which integrate predictor optimization into the selection process, and filter models, which select features based on the overarching characteristics of the training data without dependency on any predictor (Chen et al., 2015). For instance, correlation-based feature selection, a filter technique, prioritizes feature subsets using a correlation-based heuristic evaluation function. Although filter methods are generally more computationally efficient, wrapper models often yield superior outcomes.

In the context of forecasting regional financial performance, particularly through sub-sovereign credit ratings, the terms "wrappers" and "filters" denote distinct methodologies for feature selection. This process is pivotal in modeling, as it entails selecting a subset of pertinent features (variables or predictors) to enhance model efficacy. Wrapper methods assess feature subsets by employing different combinations in model training and testing. The model's performance with a specific learning algorithm serves as the benchmark for identifying the optimal feature subset. This iterative approach involves evaluating various feature combinations based on model performance, with the wrapper method favoring features that significantly enhance model prediction within a given context. It is adept at identifying feature interdependencies tailored to the model, although it may be computationally intensive for extensive feature sets and prone to overfitting with limited data. On the other hand, wrapper methods are particularly valuable when feature interactions are crucial for accurate predictions, potentially elevating model performance at the expense of increased computational demands. Such methods are adept at discerning subtle dynamics among predictors of creditworthiness.

The illustration below (Figure 3) depicts the algorithm of the wrapper method.

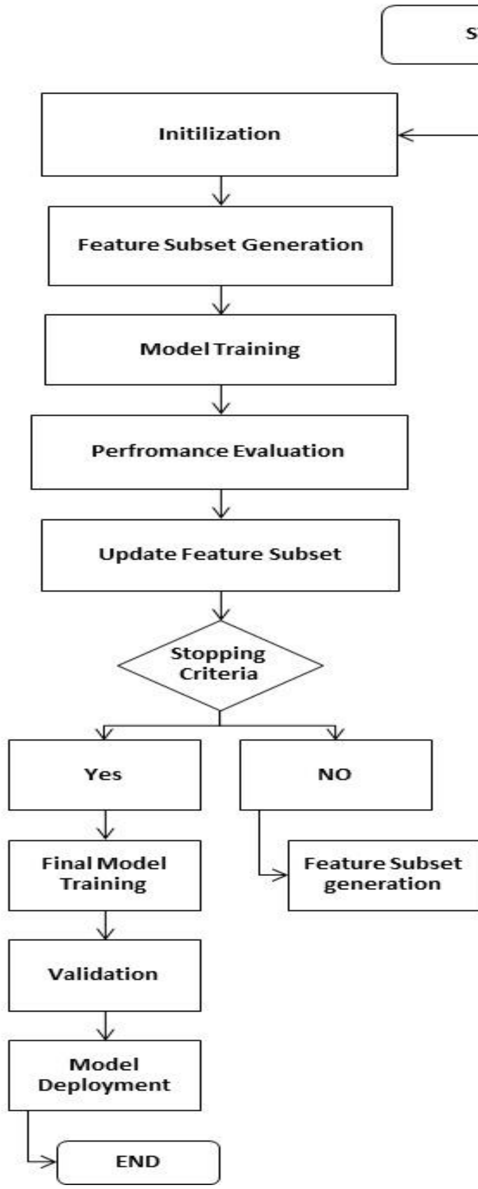
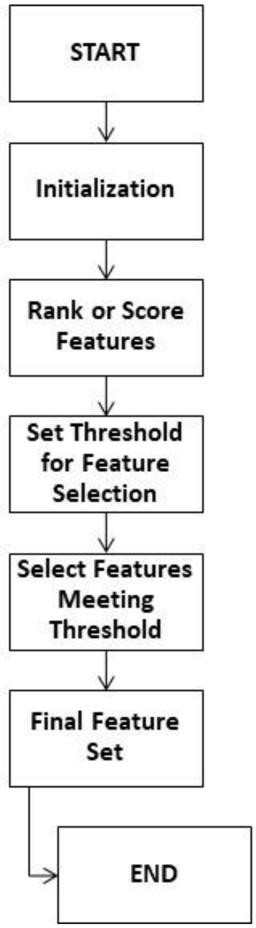


Figure 3: Algorithm of the wrapper feature selection method

Filter methods, conversely, are suited for large datasets and scenarios with constrained computational resources, focusing on the individual relevance of features rather than their interplay. They offer a rapid assessment of feature significance, providing an efficient means to gauge the importance of each predictor independently.

Filter methods employ variable ranking techniques as the primary criteria for variable selection by ordering. These methods are used due to their simplicity and reported success in practical applications. A suitable ranking criterion is used to score the variables, and a threshold is applied to remove variables below the threshold. Ranking methods are considered filter methods since they are applied before classification to filter out less relevant variables. A fundamental characteristic of a distinctive feature is to contain valuable information about the various classes in the data. This characteristic can be defined as feature relevance, which measures the usefulness of the feature in distinguishing the different classes (Chandrashekar et al., 2014).

Filtering methods evaluate the relevance of features independently of a specific learning algorithm. Features are scored based on certain criteria, e.g., correlation, mutual information, and a subset is selected based on features that are ranked or scored according to a certain criterion. A threshold is set and features that meet the criterion are selected. Computational filter methods are very efficient and fast compared to wrapper methods and are less prone to overfitting because they do not involve iterative training of the model and provide insight into the importance of features. As a result, interactions between features that contribute to predictive power are lost. The features selected may not be optimal for a particular learning algorithm. However, the filter method can be used for a quick feature selection process and the wrapper method for fine-tuning and capturing interactions. The choice between wrapper and filter methods depends on the specific context, the size of the dataset and the importance of the feature interactions. Figure 4 below depicts the algorithm of the filter methods.



.

Figure 4: Algorithm of filter feature selection methods

3.3 Classification Methods

This section discusses important soft computing methods, specifically ensemble learning and single classification methods that can serve as base classifiers in ensembles and were chosen to predict regional financial performance due to their success in previous studies on corporate credit rating prediction (Wang and Ku, 2021; Luo, 2022; Yu et al., 2022).

3.3.1 Statistical models

Logistic regression estimates probabilities through a logistic function, which represents the cumulative distribution function of logistic distribution. This model follows a similar approach to probit regression, which uses a cumulative normal distribution curve instead. Linear regression and logistic regression are both statistical models used to measure the association between dependent and independent variables:

$$\left[\frac{P_i}{1-P_i}\right] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \quad (1)$$

The formula for calculating the probability of the outcome is denoted by P_i, where subscript i represents the i-th observation in the sample. The intercept term is symbolized by β_0, and the coefficients linked to each explanatory variable $X_1, X_2, \ldots, X_k$ are represented by $\beta_1, \beta_2, \ldots, \beta_k$.

Ordered Probit Model

The ordered probit model is commonly used for regression analysis, particularly for dealing with qualitative ordinal dependent variables, such as ratings that are classified into more than two categories. This model expresses the probability of credit ratings assigned to a country as a function of a set of explanatory variables. The dependent continuous variable that measures creditworthiness, Y, is a linear function of a set of explanatory variables X, with a parameter vector β, and a term:

$$prob(Y_i = 1|X_i) = \int_{-\infty}^{x'\beta} \emptyset(t)dt = \emptyset(X'i\beta) \quad (2)$$

- β is a vector of parameter estimates,
- $\emptyset$ is a cumulative distribution function (the normal, logistic, or extreme value),
- X is a vector of explanatory variables,
- P is the probability of a response,
- t is the natural (threshold) response rate.

The complementary log-log model can be expressed as log (- log [1- F (t; X)]) = $X\beta$. Here, F (t; X)]) is a cumulative distribution function (CDF) of the survival time at time t given covariates X, where

- X is the vector of the covariates.

- β *is* the vector of the coefficient.

The formula reflects the complementary log-log transformation, which introduces asymmetry into the hazard function. The left-hand side of this equation is known as the complementary log-log transformation. The complementary log-log transformation alters the response from a range of (0,1) to (-∞, +∞). Unlike the logit and probit, the complementary log-log model displays asymmetry and is commonly applied when the event's probability is significant:

$$ln(-ln(1 - P(Y = 1)) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k \quad (3)$$

where:

- $P(T < t)$ is the probability that the event of interest (failure or survival) occurs by time t,
- $\beta_{0,} \beta_1, \ldots, \beta_k$ are the coefficient of the model,
- $x_1, x_2, \ldots, x_k$ are the predictor variables.

In this formula, the complementary log-log function is applied to the linear predictors. The exponential function and the negative sign are used to transform the linear predictor in a way that suits survival analysis.

Cauchit model calculates the transformation for the Cauchit or tangent link, its inverse, and the first two derivatives. This link function provides an alternate way of fixing parameters that fall within the unit interval. The relationship this link has to the Cauchy distribution is like that of the probit link to Gaussian. A noteworthy characteristic of this link function is its heavier tail when compared to other link functions. Numerical values of f that are close to 0 or 1 should be noted:

$$f(x; x_0; \vartheta) = \frac{1}{\pi\vartheta \left[1 + (\frac{x - x_0}{\vartheta})^2\right]} \quad (4)$$

where ϑ is the scale parameter, which controls the width of the distribution, and x_0 is the location parameter, which specifies the location of the peak of the distribution.

3.3.2 Single classifiers

There are numerous single classifiers in machine learning, each designed to address specific types of classification problems. Below is a list of some commonly used single classifiers:

Decision trees are non-parametric supervised learning models used for classification and regression tasks. They recursively partition the feature space into regions, with each partition associated with a specific decision or outcome. At each node of the tree, a decision is made based on the value of a selected feature, resulting in a partition criterion that optimally separates the data according to certain criteria, often maximizing information gain or minimizing impurity. Decision trees are interpretable, making them valuable for understanding the decision process and identifying important features in the data. However, they are prone to overfitting, especially with deep trees, which can be mitigated by techniques such as pruning, setting a minimum number of samples required to split a node, or using ensemble methods such as random forests. Despite this limitation, decision trees are still widely used because of their simplicity, interpretability and flexibility in handling both categorical and numerical data.

The Naive Bayes algorithm is a probabilistic classifier based on Bayes' theorem, which assumes that features are conditionally independent given the class label, hence the "naive" assumption. It calculates the probability of a class label given a set of features by multiplying the conditional probability of each feature given the class label and the prior probability of the class label itself, and then normalizing to obtain a probability distribution over the possible class labels. Despite its simplistic assumption, Naïve Bayes often performs remarkably well in practice, especially for text classification tasks, due to its efficiency, ease of implementation, and ability to handle high-dimensional data with sparse feature sets. However, its performance can suffer when the independence assumption is violated or when features are highly correlated. Nevertheless, Naïve Bayes remains a popular choice for classification tasks, particularly in situations where computational resources are limited or interpretability is critical. A Bayesian network, or Bayes net, is a type of probabilistic graphical model that represents a set of random variables and their conditional dependencies through a directed acyclic graph. This provides a transparent and modular model equipped with probabilistic reasoning. However, it is necessary to specify the model, and as the number of variables increases, so does the computational complexity.

The SVM algorithm is a powerful supervised learning method, primarily used for classification tasks, but also extended to regression and outlier detection. SVMs work by finding the optimal hyperplane that separates data points of different classes in a high-dimensional space, maximizing the margin between classes. This hyperplane is determined by support vectors, which are the data points closest to the decision boundary. SVMs use a kernel function to map the input data into a higher-dimensional feature space where non-linear decision boundaries can be represented linearly. This flexibility allows SVMs to effectively handle complex data distributions and achieve high generalization performance. However, SVMs can be sensitive to the choice of kernel and parameters, and training large datasets with SVMs can be computationally intensive. Nevertheless, SVMs are widely used in various fields due to their ability to handle both linear and non-linear classification problems and their strong theoretical foundations. The sequential minimal optimization algorithm is an iterative method used to solve quadratic programming problems and can therefore be used to train SVM.

The MLP classifier is a versatile NN architecture commonly used for classification tasks. It consists of multiple layers of nodes, including an input layer, one or more hidden layers, and an output layer. Each node in the network applies a non-linear activation function to the weighted sum of its inputs, allowing the MLP to model complex, non-linear relationships in the data. During training, the network learns to adjust weights and biases through backpropagation, minimizing a predefined loss function such as cross-entropy. MLPs are highly flexible and can approximate any continuous function, given sufficient data and computational resources. However, they are prone to overfitting, especially with large, complex data sets, requiring regularization techniques such as dropout or L2 regularization. Despite their complexity, MLPs have been successfully applied in several domains, including credit rating prediction (Overes and van der Wel, 2023).

3.3.3 Ensemble classification methods

The primary objective of an ensemble classifier is to decrease the misclassification rate or error rate of a weak classifier by combining multiple classifiers. This is achieved by obtaining predictions from several classifiers on the original data and merging them to create a robust classifier. Enlisted below are some examples of ensemble classifiers.

Bagging (Breiman, 1996), also known as Bootstrap Aggregating, is a technique used to reduce the variance of forecasts. It does this by creating supplementary datasets for training novel datasets using combinations with recurrences to produce multisets of the same size as the novel data. Bagging is used to generate multiple forms of prediction. When the size of the training set is increased, it does not necessarily improve the predictive power. However, reducing the variance can help to bring the prediction closer to the expected result. In such cases, the new data set can be used to combine predictors. The ensemble method offers an alternative approach by creating artificial training data, which increases the diversity of base classifiers. This makes it particularly suitable for smaller datasets.

Algorithm 1: Bagging

Input	The set T of training data (x^i; y^i), i=1,2, … ,n; the number B of base classifiers
Output	Ensemble of base classifiers $\{C_b\}$
For b=1 to B	Create a bootstrapped replicate T_b of the training data set T; Construct a base classifier C_b on T_b;} Combine base classifiers C_b, b=1,2,…,B into an ensemble $\{C_b\}$ by simple majority voting;

Random forest is a highly popular and powerful machine learning algorithm used for financial performance forecasting (Yeh et al., 2012). The algorithm employs a small twist to bagging by exploiting random feature selection. It is an ensemble learning method used for classification, regression, and other tasks that operate by building a multitude of decision trees. Random forest is a supervised machine learning algorithm that combines numerous classifiers to enhance model performance. It is composed of decision trees.

Algorithm 2: Random forest

Input	The set T of training data (x^i; y^i), i =1,2,….., n; the number B of base classifiers
Output	Ensemble of base classifiers $\{C_b\}$
For b=1 to B	Bagging + randomly select % of possible splitting features N\{node in the tree\}, select feature X_j with the highest information gain to split on from the original training data set T; Construct a base classifier C_b in T_b;
Create nodes	N_1... N_j, where X_j has the values $X_{j1 \ldots} X_{jn}$

AdaBoost (Freund et al., 1996; Koutanaei et al., 2015) can be used to improve the performance of any machine learning algorithm such as decision trees by sequentially improving the performance of base classifiers. These classifiers are simple and contain only one decision for classification and are therefore called decision stumps. Each instance in the training dataset is weighted. The initial weight is set to weight(x) = 1/n, where x is the training example, and n is the number of training examples. AdaBoost is best used to boost the performance of decision trees by placing greater weight on examples misclassified by the preceding classifiers.

Algorithm 3: AdaBoost

Input	The set T of training data (x^i; y^i), i=1,2, … ,n; the number B of base classifiers
Output	Ensemble of base classifiers $\{C_b\}$
For b=1 to B	Construct a base classifier C_b on weighted training data $T^*=(w_1T^1{}_b, w_2T^2{}_b, \dots, w_nT^n{}_b)$; Calculate the probability estimates of the error $\mathrm{err}_b=1/n\ \Sigma\ w_{ib}\times\xi^i{}_b$ ($\xi^i{}_b=0$ if T^i classified correctly, $\xi^i{}_b=1$ otherwise); Set weight $c_b=0.5\times\log((1-\mathrm{err}_b)/\mathrm{err}_b)$; If $\mathrm{err}_b<0.5$, set $w_{ib+1}=w_{ib}\times\exp(c_b\xi^i{}_b)$; Otherwise, set all weights $w_{ib}=1$ and restart the algorithm;}
Combine base classifiers	C_b, b=1,2,…,B into an ensemble $\{C_b\}$ by weighted majority voting;

The LogitBoost algorithm (Friedman et al., 2000) is a backfitting additive logistic classifier. It justifies the exponential bound established in the preceding section as an estimate to the objective function derived when a generalized additive linear model is used to fit a classification problem after certain logistic transform. Logitboost attempts to minimize the likelihood of the classifier, which in turn is restricted to a parametric family of density functions. LogitBoost performs additive logistic regression and, thus, can effectively handle multi-class problems.

XGBoost, short for extreme Gradient Boosting, is a boosting algorithm that sequentially trains weak models, which are decision trees, on the same data. Chen and Guestrin (2016) introduced XGBoost. It uses gradient descent optimization to minimize the loss function and a regularization technique to prevent overfitting. The regularization technique involves adding a penalty term to the loss function.

Stacking, as described by Tsai and Chen (2010), involves training a learning algorithm to combine the predictions of multiple other learning algorithms. The individual algorithms are trained using available data, and then a combiner algorithm is trained to make a final prediction using all the predictions of the other algorithms as additional inputs. When using an arbitrary combiner algorithm, stacking can theoretically represent any of the ensemble techniques described in this article. However, in practice, a single-layer logistic regression model is often used as the combiner.

Voting ensembles (Finlay, 2011) can be used for both classification and regression problems.

It’s usually selected multiple sub-models and allow another model to specify and learn how best to combine the predictions from the sub-models. A meta model is used for the best combination of the predictions of sub-models. This technique is sometimes called blending, as it blends the predictions together. The built classifier is selected from a set of classifiers based on the error on the training data, instances serve as training data.

DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) is a machine learning algorithm that aims to reduce overfitting in ensembles by using additional artificially generated training data and introducing diverse training examples through oppositional relabeling. The algorithm achieves this by adding different randomly constructed instances to the training set, generating highly diverse ensembles. This helps to make the base classifiers less prone to memorizing noise in the data (Patel et al., 2013).

3.3.4 Rule-based classification methods

JRip algorithm, also known as JRip (J48 Rules), is a rule-based classification algorithm which implements a propositional rule learner (Repeated incremental Pruning to produce Error Reduction (RIPPER)). It is an extension of the C4.5 decision tree algorithm and is designed to extract classification rules from a dataset. JRip was introduced as part of the Weka machine learning software, which is a popular open-source collection of machine learning algorithms. The algorithm begins by constructing a decision tree using the C4.5 algorithm. The decision tree is a collection of rules that can be derived from the dataset. Each path from the root to a leaf node corresponds to a rule. Once the decision tree is constructed, JRip extracts a set of rules from it. Each rule consists of a combination of conditions on the data features that lead to a specific class prediction. JRip applies a pruning process to simplify and improve the interpretability of the rules. This process removes conditions from rules that do not significantly contribute to their accuracy. The algorithm optimizes the extracted rules to improve their compactness and reduce redundancy in rule optimization. This step involves combining or simplifying rules without sacrificing their accuracy (Veeralakshmi, 2015).

Fuzzy rule-based systems begin by defining a set of input variables and their linguistic terms, represented as fuzzy sets that capture the data's uncertainty. The evolutionary scheme can approximate and describe fuzzy rule bases. As the evolutionary algorithm progresses, the population of rule sets evolves to better fit the predicted classes, and the best rule set is selected as the final model. The Fuzzy Unordered Rule Induction Algorithm (FURIA) (Huhn and Hullermeier, 2009) is a machine learning method specifically designed for rule-based classification tasks. FURIA operates by generating fuzzy if-then rules from the dataset, where each rule consists of conditions and corresponding class labels represented in fuzzy logic terms. It employs a heuristic approach to iteratively refine and optimize these rules, aiming to maximize classification accuracy while minimizing rule complexity. FURIA considers the unordered nature of features, allowing it to handle datasets with categorical or nominal attributes efficiently. By utilizing fuzzy logic, FURIA can capture the uncertainty and vagueness inherent in real-world data, enabling robust classification in domains where traditional crisp rules may not suffice. FURIA is particularly beneficial for datasets with noisy or imprecise features, providing a flexible and interpretable framework for rule-based classification tasks in various fields, including pattern recognition, decision support systems, and expert systems.

FURIA is an extension of the well-known RIPPER algorithm, which is a state-of-the-art rule learner, while maintaining its superiority, such as simple and coherent rule sets. This method allows for a more precise and flexible classification method. FURIA learns fuzzy rules rather than conventional rules and unordered rule sets instead of rule lists. Moreover, it employs an effective rule extension method to handle uncovered examples.

3.3.5 Cost-sensitive ensemble classification

In the thesis, I employed the use of the MetaCost classifier, a principled method for making an arbitrary classifier cost-sensitive by wrapping a cost-minimizing technique around it (Toseafa and Hajek, 2019). This technique treats the fundamental classifier as a black box, requiring no knowledge of its functioning or a change to it. MetaCost is applicable to any number of classes and to arbitrary cost matrices. Realistic trials on a large group of target databases show that MetaCost almost always yields large cost cutbacks compared to the cost-blind classifier used. The MetaCost operator makes its base classifier cost-sensitive by using the cost matrix specified. MetaCost relies on an internal cost-sensitive classifier in order to relabel classes of training examples, see Domingos (1990). The MetaCost operator is a nested operator and has a sub-process. The sub-process must have a learner as the operator that expects an example set to generate a model. This operator tries to build a better model using the learner provided in its sub-process. MetaCost is based on the Bayes optimal prediction that reduces the expected cost $R(j|x)$ (Michie et al., 1994):

$$R(j|x) = \sum_{i}^{I} p((j|x)cost(i,j)) \tag{5}$$

where $p(j|x)$ is the probability of class j given example x and cost(i,j) is the cost of misclassifying a class i example as class j. The Bayes optimal prediction rule implies a division of the example space into I classes, such that class i is the minimum expected cost prediction in class i. If misclassifying class i becomes more expensive relative to misclassifying others, then parts of the former non-class i regions shall be re-allocated as class i since it is now the minimum expected cost prediction.

3.4 SHAP Values

The Shapley (SHAP) value, named after Lloyd Shapley, who introduced it in 1951 and won the Nobel Memorial Prize in Economic Sciences for it in 2012, is a solution concept in cooperative game theory. Furthermore, SHAP is a mathematical method used to explain machine learning model predictions. It is based on game theory concepts and can calculate the contribution of each feature of the prediction. This allows for the determination of the most important features and their influence on the model's prediction.

The Shapley value is a way to fairly distribute the gains of a cooperative game among the players. The game is defined by a set of players and a payoff function, which assigns a payoff to each player for each possible combination of players. The Shapley value is calculated by considering all possible combinations of players and the marginal contribution of each player to the payoff of the coalition. This value has several desirable properties, including efficiency, symmetry, linearity, and the null player property. Efficiency means that the sum of the Shapley values of all players is equal to the total payoff of the game. Symmetry dictates that players who have an equal impact on the payoffs of all coalitions will receive the same Shapley value. Linearity states that a player's Shapley value is equal to the sum of their Shapley values in all possible subgames. The null player property ensures that a player who has no impact on the payoffs of any coalition will receive a Shapley value of zero.

The Shapley value has several applications in economics, finance, and machine learning. For example, it can be used to fairly distribute the profits of a cooperative venture, to allocate resources among different departments in a company, or to identify the most important features in a machine learning model (Lundberg and Lee, 2017).

4 Experimental Settings

4.1 Hardware and Software Specification

The experiments were conducted on a PC equipped with a Core i5 processor that operates at a clock speed of 2.4 GHz, complemented by 16 GB of central RAM. For the operating system, Windows 10 (64-bit) was selected as the platform of choice.

Regarding software specification, experiments were carried out in Weka 3.8.3 x 64 program environment (data preprocessing, feature selection, and single classifiers, including FURIA), STATISTICA 13.1 (statistical methods, including various versions of logistic regression), and scikit-learn (XGBoost) and SHAP libraries in Python 3.12 environments.

Weka (Waikato Environment for Knowledge Analysis) was chosen as a popular machine learning software written in Java and developed at the University of Waikato, New Zealand. It stands out from other machine learning platforms for several reasons. Weka has a graphical user interface that is intuitive and easy to use, allowing for rapid visualization, analysis and interpretation of data and results. Weka also comes with a wide range of tools for data preprocessing and classification algorithms. As open-source software, Weka allows its source code to be viewed, modified and distributed, facilitating research and development in machine learning. While Weka has many advantages, it is important to note that it may not be the best fit for very large datasets or highly specialized machine learning tasks that require the use of deep learning frameworks such as TensorFlow or PyTorch. However, for standard machine learning tasks, such as those addressed in this thesis, Weka provides an easy-to-use and comprehensive platform.

STATISTICA 13.1 is a program for data analysis and visualization[2]. It offers a wide range of algorithms, functions, tests, and methods for data analysis, including simple breakdown tables, advanced nonlinear modeling, generalized linear models, and time-series methods.

Python 3.12, an advanced and high-level programming language, is interpreted and object-

[2] https://statistica.software.informer.com/

oriented with dynamic semantics. Its inherent data structures, along with dynamic typing and binding, enhance its suitability for rapid application development, as well as for scripting or bridging different components. The simplicity and clarity of Python's syntax makes it easy to learn, promotes readability and minimizes maintenance. It also supports modularity and code reuse through its robust system of modules and packages. Available free of charge in both source and binary formats for all major platforms, Python's interpreter and comprehensive standard library can be freely distributed, further contributing to its widespread adoption and versatility.

4.2 Data Preprocessing and Partitioning

The first step was to replace missing values with mean values. This is a widely used imputation method in classification tasks due to its simplicity, its computational efficiency, and its ability to preserve the integrity of the data without discarding any valuable information. This approach helps to preserve dataset size, reduce potential biases associated with systematic missing data, and ensure compatibility with machine learning algorithms that require complete datasets. Mean imputation serves as a practical baseline. However, it can lead to reduced variance and potentially biased estimates, particularly in datasets with non-normally distributed features or a high proportion of missing values, where more sophisticated imputation methods may be preferable. In the case of the Moody's data, however, this was not an issue, with less than 1% of data missing.

Second, the data were normalized to the [0,1] scale, as this is critical in classification tasks to ensure that features scale uniformly, improving model convergence, accuracy and interpretability. It prevents any single feature, particularly in algorithms sensitive to feature scale such as *k*-NN and SVM, from disproportionately influencing the model's decisions. In addition, normalization facilitates faster convergence in gradient descent-based algorithms, ensures numerical stability, and fulfils the assumptions of certain models that expect data to be on a similar scale, by bringing all features to a similar scale. This preprocessing step helps both the optimization process and the prevention of overfitting and contributes to a more balanced and effective classification model.

Data partitioning in machine learning refers to the process of dividing a dataset into two or more subsets, typically for the purpose of training and testing machine learning models. The primary goal of data partitioning is to evaluate the performance of a model on unseen data, thereby assessing its ability to generalize to new, unseen instances. The most common types of data partitioning consist of the training dataset, which is the part of the dataset used to train the classification model (machine learning model). The model learns patterns, relationships, and features from this set. Datasets are usually split into training and testing subsets in order to benchmark the algorithm performance. One model is trained on the training dataset, the testing dataset is used to evaluate it. The split ratio for data partitioning is commonly expressed as a percentage of the total dataset. For this thesis, a 67-33 split was used, implying that 67% of the data is used for training, while the remaining 33% is used for testing the model (Sarkar, 2016). However, performing the process only once and randomly has a serious constraint and can lead to sample selection bias to resolve this issue. To obtain a reliable estimate of classification performance, this process was stratified and repeated five times.

Furthermore, 10-fold cross-validation was used to find the optimal settings of the hyperparameters during the training process of the classification models. For *K*-fold cross-validation, this approach involves randomly splitting the dataset into *K* equally sized parts and training the model *K* times. In each training cycle, a single partition is selected that has not been chosen in the previous cycles. The selected portion is used for testing, while the remaining dataset is used for training. As a result, each model will be trained and tested on a distinct training dataset. After all *K* cycles have been completed, the results are aggregated. Research indicates that setting 10 as the *K* value produces dependable results, avoiding both excessively high bias and variance (Kohavi, 1995).

For feature selection, the correlation-based filter and a wrapper were used. For both algorithms, the space of feature subsets was searched using greedy hillclimbing augmented with a backtracking facility. The algorithms started with the empty set of features and searched the space forward. Area under ROC curve was used as the classification objective criterion due to the imbalance in the data.

4.3 Settings of Classification Methods

The *k*-NN classifier was implemented with the Euclidean distance function and *k* set to 3. This thesis employed the K2 algorithm to explore the search space of Bayes net with a simple estimator.

SVM was trained using the sequential minimal optimization (SMO) algorithm with complexity parameter $C = \{1, 2, \dots, 128\}$, polynomial kernel function with exponent = {1, 2}, and RBF kernel function with gamma = 0.01. Its complexity parameter was determined using the grid search procedure, ranging from 1 to 128.

The NN model was trained using backpropagation with the following settings: hidden layer neurons ranged from 1 to 100 (the grid search procedure used to find the optimal number), the learning rate was set to 0.1, and the number of iterations was set to 500 (Overes and van der Wel, 2023).

The J48 algorithm, developed by Ross Quinlan, was used to generate a decision tree. The J48 version utilized a minimum of two instances per leaf and a confidence factor of 0.25 for pruning. The J48 decision tree algorithm also served as the base learner in the Decorate ensemble method, with 15 classifiers in the ensemble and an ensemble diversity of 1.0. For the random forest model, 100 trees were generated with unlimited maximum depth and the number of variables randomly sampled as candidates at each split was calculated as log2(#predictors) + 1. The AdaBoost M1 version was trained with decision stump as base learners, and the number of iterations was 10. Bagging was trained with REPTree as the base learner. LogitBoost was trained with a decision stump as the base learner, and the Z max threshold for responses was set to 3. XGBoost was trained with gbtree as booster, learning rate of 0.3, minimum loss reduction of 0, maximum depth of a tree set to 6, minimum sum of instance weight (hessian) needed in a child of 1, L1 regularization term on weights of 0, and L2 regularization term on weights of 1. Finally, Voting was performed using an ensemble of classifiers including Bayes net, Naïve Bayes, SVM, random forest, and J48 decision tree.

For JRip, the minimum total weight of the instances in a rule was set to 2.0, the number of optimization runs was 2, and pruning was enabled. The same values of hyperparameters were also used for its fuzzy counterpart FURIA. In addition, fuzzy partitions were enabled, and rule smoothing was enabled.

4.4 Evaluation Measures

While evaluating the experimental results, the following evaluation measures were taken into consideration: Accuracy (Acc), AUC (to evaluate the performance on imbalanced classes), Precision, Recall, F1 measure (the combination of precision and recall), False positive rate (FPR), Kappa, Matthews correlation coefficient (MCC), and Misclassification cost (to consider different financial effects of misclassified regional units).

Accuracy is a commonly used metric in machine learning to evaluate the performance of a classification model. It is a measure of the overall correctness of the model in predicting the class labels of the instances in the dataset. Accuracy of a classification model is calculated as the ratio of correctly predicted instances to the total number of instances in the dataset. Accuracy is expressed as a percentage, ranging from 0% to 100%. A higher accuracy indicates a better-performing model. Accuracy (Acc) is the percentage of classes which were predicted correctly:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

where *TP*, *TN*, *FP*, and *FN* are the numbers of true positives, true negatives, false positives, and false negatives.

TPR and FPR are metrics commonly used to evaluate the performance of binary classification systems. TPR, also known as sensitivity or recall, measures the proportion of actual positives that are correctly identified by the classifier:

$$TPR\ (Recall) = \frac{TP}{TP + FN} \tag{7}$$

The FPR measures the proportion of actual negatives that are incorrectly classified as positives by the classifier:

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

Precision measures the accuracy of positive predictions made by the classification model. It is defined as the number of true positive predictions divided by the total number of positive predictions:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

The F1 measure is the harmonic mean of Precision and Recall. It combines Precision and Recall into a single metric by taking their harmonic mean to give an overall effectiveness of the classification model in terms of both false positives and false negatives:

$$F1\ measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

MCC takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes:

$$MCC\ = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(T + FP)(TN + FN)}} \tag{11}$$

Kappa, or Cohen's Kappa, is a statistical measure used to evaluate the reliability of agreement between two raters (or sets of data) that classify items into mutually exclusive categories. It is particularly useful when assessing the degree of agreement beyond what would be expected by chance. A Kappa value of 1 indicates perfect agreement between the raters, while a Kappa value of 0 indicates that the agreement is no better than what would be expected by chance.

The ROC curve is a graphical representation of the performance of a classification model at different settings of the threshold value. It is a probability curve, and the AUC represents the degree of separability, or the two-dimensional area underneath the entire ROC curve. This indicates how well the model can distinguish between classes. The AUC is a measure of how well the model predicts 0 classes as 0 and 1 classes as 1. It represents the probability that the classifier will rank a randomly chosen class higher than a class.

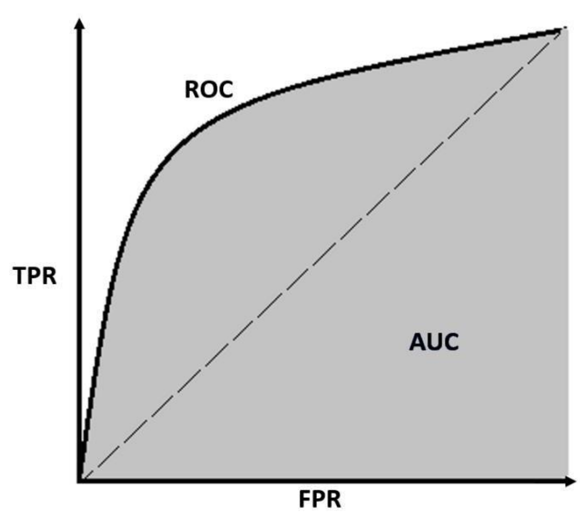


Figure 5: ROC curve and AUC

In the literature on credit rating modelling, AUC was reported to be a suitable performance measure, mainly because it is robust against imbalanced data:

$$AUC = \int_0^1 TPR(T) \times \frac{d}{dT} FPR(T) dT, \tag{9}$$

where T is any cut-off point, $0 < T < 1$. On the one hand, the wrong prediction of a class that is (type II error) leads to the loss of time.

To calculate the overall Misclassification cost, the cost matrix is proposed as presented in Table 5.

Table 5: Cost matrix for eight rating classes Aaa to Ca

	Rating class target (actual)					
Rating class predicted	Aaa	Aa	A	Baa	…	Ca
Aaa	0	1	2	3	…	8
Aa	1	0	1	2	…	7
A	2	1	0	1	…	6
Baa	3	2	1	0	…	5
…	…	…	…	…	…	…
Ca	8	7	6	5	…	0

Note: only 8 classes were used because the remaining classes were not present in the data

5 Experimental Results

In this section, I present the results of the empirical experiments conducted to evaluate the effectiveness of the proposed classification models on the two benchmark datasets from Moody's credit rating agency. Subsequently, the averages and standard deviations of the five stratified train/test of 67/37% splits are presented. For each classification method, the best and statistically similar results at $P < 0.05$ are presented using the Wilcoxon signed-rank test, which is a non-parametric statistical hypothesis test used when comparing two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ. The advantages of using the Wilcoxon signed-rank test for this purpose are as follows: a) the Wilcoxon test does not require the assumption of normality, b) the Wilcoxon test is less sensitive to outliers than parametric tests, and c) the Wilcoxon signed-rank test can be used with relatively small sample sizes.

5.1 Performance of Logistic Regression Models

Table 6 compares the performance of different link functions used in statistical models, such as Complementary Log-Log, Probit, Logit, and Cauchit, for the benchmark dataset with 2016 credit rating outcome. The table provides metrics of model fit, goodness of fit, and a measure of explanatory power (pseudo R-squared). It is worth noting that each link function has a different way of relating the linear predictor to the mean of the distribution function, with Complementary log-log often used for models where the outcome is time to an event (e.g., survival models), while Probit is usually applied when the outcome is binary, as in probit regression, Logit (in logistic regression) is perhaps the most popular link function for binary data, and Cauchy can be useful when outcomes are with heavier tails than the logit model assumes.

The values of model fitting/chi-square indicate how well the model fits the data, with higher values generally indicating a better fit. It is based on the chi-square test statistic, which compares the observed values with the values expected by the model. The Complementary log-log model has a chi-square value of 899.031, indicating a relatively good fit compared to the other models. The fit of the Probit model is also better than the Logit and Cauchit models, with a chi-square value of 361.696. The Logit model has the lowest chi-square value (242.188), suggesting that it may be the least well-fitting of the models.

Unlike the fitting/chi-square, higher values goodness-of-fit/Pearson indicate a poorer fit, as they suggest a greater discrepancy between observed and expected frequencies. The Complementary log-log model has a relatively low Pearson value (1108.241), indicating a good fit. The Probit and Cauchit models have higher Pearson values, indicating a poorer fit. The Logit model has the highest Pearson value, again suggesting that it may not fit the data as well as the other models.

The Pseudo R-Square/Cox and Shell metric provides an estimate of the proportion of variance in the dependent variable that is explained by the model. Higher values indicate that the model explains more variance in the response variable, which is desirable. Again, the Complementary log-log model has the highest pseudo R-squared value (0.972), indicating that it explains the variance in the response variable very well. The explanatory power of the Probit model is substantially lower (0.763), followed by the Cauchit (0.674) and Logit (0.619) models.

Overall, the Complementary log-log model appears to perform best of all the metrics presented, indicating a good fit to the data and high explanatory power. The Probit model also shows a good fit but has less explanatory power. The Logit and Cauchit models appear to perform less well in terms of both fit and explanatory power.

Table 6: Results of different link functions

Link function	Model fitting/Chi-Square	Goodness-of-fit/Pearson	Pseudo R-Square/ Cox and Shell
Complementary Log–log	899.031	1108.241	0.972
Probit	361.696	3007.563	0.763
Logit	242.188	4253.489	0.619
Cauchit	281.337	4180.606	0.674

Table 7 below presents the significant results empirically. The analysis shows that credit ratings are not significantly influenced by financial indicators such as GDP per capita and real GDP. This finding is surprising given that previous studies have shown that prevailing economic conditions in countries are a significant factor affecting credit ratings (Afonso et al., 2011; Boumparis et al., 2019). A region's high GDP may not necessarily lead to a positive credit rating score due to factors such as large budget deficits and excessive accumulated debt. The study also found that national unemployment is a statistically significant determinant of credit rating in the sampled regions at a 95% significance level. The creditworthiness of regions is significantly influenced by the level of economic stability and the general economic environment. This finding is consistent with Afonso et al.'s (2011) research, which also found a positive association. Additionally, this work found that neither short-term nor long-term debt variables were significant factors in determining credit rating. However, the thesis found that debt maturity is negatively related to credit rating, indicating that it reduces the likelihood of a good rating. This can be explained by the fact that a country's inability to refinance its maturing debt could lead to financial stress, potentially resulting in a credit rating downgrade. This result is consistent with the findings of Sajjad and Zakaria (2018), who also found a negative association. The variables OwnRev/OperRev, GovTransf/OperRev, EarRev/OperRev, and Inter/OperRev were statistically insignificant factors driving credit ratings. The results indicate that only DebtSer/TR was found to be negatively associated with credit ratings. Additionally, the variables CashSurp/TR and TEPerCapita were found to be significant but negatively related to credit ratings. Refer to Appendices A to C for all the results from the experiments performed for the other link functions.

Table 7: Parameter estimates for the Complementary log–log model

Variables	Estimate	Standard Error	Wald Stat.	*P*-value
Threshold [Aaa = 1]	-2.629	0.537	23.942	0.000***
[Aa = 2]	-1.235	0.483	6.539	0.011**
[A = 3]	-0.530	0.471	1.263	0.261
[Baa = 4]	0.126	0.467	0.073	0.787
[Ba = 5]	1.254	0.471	7.079	0.008**
[B = 6]	2.128	0.489	18.927	0.000***
[Caa = 7]	2.558	0.514	24.795	0.000***
GDPPerCapita	-7.447E-5	0.001	0.003	0.956
RealGDP	-0.042	0.046	0.827	0.363
Unemployment	- 0.031	0.023	1.739	0.187
NationalUnemployment	0.073	0.025	8.577	0.005**
Debt/GDP	-0.004	0.009	0.220	0.639
FXDirectDebt(beforeswap)	0.016	0.005	8.226	0.005**
ShortDebt/Debt	0.011	0.007	2.779	0.095
LongDebt/Debt	0.002	0.002	0.768	0.381
MaturityDebt	-0.066	0.018	13.940	0.000***
OwnRev/OperRev	-0.007	0.004	3.285	0.070
GovTransf/OperRev	0.003	0.004	0.647	0.421
EarRev/OperRev	-0.001	0.005	0.031	0.861
Inter/OperRev	0.087	0.064	1.846	0.174
DebtSer/TR	-0.030	0.015	3.809	0.051*
AccualFinancingSurplus/TR	0.003	0.020	0.021	0.884
CashSurp/TR	-0.035	0.018	3.856	0.050*
TEPerCapita	-8.438E-5	2.841E-5	8.825	0.005**
TE/GDP	0.017	0.015	1.248	0.264
OperBalance/OR	-0.007	0.014	0.239	0.625
NetOperatingBalance/OR	- 0.008	0.008	1.043	0.307
SelfFinRatio	0.316	0.143	4.898	0.027*
CapitalSpend	0.008	0.010	0.605	0.437
NWC/TE	-0.011	0.003	9.456	0.005**

Note: The Ca class was represented with only one sample, statistically significant at * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$

5.2 Performance of Single Classifiers

Table 8 presents the performance of single classifiers in an experiment for the benchmark dataset with rating classes from 2016, as assigned by the Moody's CRA. This table shows the performance metrics of different classifiers, including decision trees (DT), neural networks (NN), support vector machines (SVM), naïve Bayes (NB), Bayesian network (BN), JRip (JRIP) and FURIA. The metrics include FPR, Precision, Recall, F1 measure, MCC, AUC, Accuracy (Acc) and Cohen's Kappa (Kappa). Values are presented as means with standard deviations (±) across five 67/33 data splits for testing data.

FPR measures the proportion of negative cases that are misclassified as positive. Lower values are better. SVM and NB have exceptionally low FPR, indicating that they rarely misclassify negative instances as positive. In contrast, higher values of Precision are desirable. NB stands out for its high precision, indicating that when it predicts an instance as positive, it is very likely to be correct. Similarly, DT and NB have relatively high recall rates, indicating that they are good at detecting positive instances. NB also scores the highest in terms of F1 measure, indicating a good balance between precision and recall. NB also has the highest MCC, indicating that its predictions have a high degree of correlation with the actual values.

Furthermore, a higher AUC indicates a better model. NB, SVM and NN have high AUC values, indicating good performance at different thresholds. While NB, NN and DT show moderate accuracies, it is important to consider this metric in conjunction with others to get a full picture of performance. NB has the highest Kappa, indicating that its predictions are in strong agreement with the actual classifications, beyond what would be expected by chance.

NB appears to perform exceptionally well across most metrics, particularly in terms of precision, F1 measure, MCC, AUC and Kappa, suggesting that it is a strong model for this dataset. SVM and NN also show strong performance, especially in terms of AUC. It is important to consider the specific application and the trade-offs between different types of error when choosing a classifier. In this scenario, false positives can be particularly costly, hence, one might prefer a classifier with a lower FPR, such as SVM or NB.

Table 8: Results of single classifiers

Classifier	FPR	Precision	Recall	F1-measure	MCC	AUC	Acc	Kappa
DT	0.065±0.030	0.319±0.188	0.560±0.339	0.408±0.239	0.399±0.216	0.763±0.173	0.506±0.048	0.416±0.061
NN	0.033±0.021	0.473±0.313	0.307±0.132	0.315±0.145	0.331±0.177	0.895±0.054	0.557±0.031	0.455±0.040
SVM	0.008±0.017	0.450±0.512	0.107±0.098	0.164±0.157	0.198±0.210	0.905±0.071	0.506±0.026	0.369±0.035
NB	0.010±0.011	0.860±0.142	0.593±0.272	0.648±0.211	0.667±0.160	0.919±0.060	0.553±0.031	0.505±0.039
BN	0.018±0.014	0.393±0.447	0.300±0.415	0.307±0.358	0.323±0.360	0.852±0.197	0.452±0.074	0.321±0.095
JRIP	0.048±0.029	0.428±0.190	0.420±0.208	0.400±0.144	0.369±0.163	0.770±0.087	0.469±0.026	0.328±0.030
FURIA	0.015±0.014	0.683±0.207	0.420±0.171	0.501±0.162	0.501±0.163	0.707±0.100	0.492±0.044	0.416±0.063

5.3 Performance of Ensemble Classifiers

Table 9 below provides performance metrics for various ensemble learning classifiers on the dataset, including bagging, random forest, AdaBoost M1, LogitBoost, and XGBoost. Random forest and LogitBoost have substantially lower FPR, indicating that they are less likely to misclassify negative instances as positive. XGBoost has relatively high Precision, suggesting that when it predicts an instance as positive, it is very likely to be correct. Random forest stands out with a significantly higher recall, suggesting that it is effective at identifying positive instances. Random forest also has the highest F1 measure, indicating a good balance between precision and recall. Random forest and XGBoost have high MCC values, suggesting that their predictions are in good agreement with the actual classifications. Random forest has a high AUC, indicating strong performance across different rating classes. Random forest has a high Kappa score, indicating a strong agreement between its predictions and the actual classifications, beyond what would be expected by chance. Finally, XGBoost has the highest accuracy, indicating that it correctly classifies a higher percentage of instances than the other classifiers.

Random forest appears to perform best in most metrics, particularly in Recall, F1 measure, MCC, AUC and Kappa, making it a strong candidate for this dataset. XGBoost also shows strong performance, especially in Precision, Acc and MCC. AdaBoost M1, while showing decent performance in AUC, has a high FPR which could be a concern taking into account the application's sensitivity to false positives. LogitBoost and bagging show moderate performance across several metrics.

Table 9: Results of ensemble classifiers

Classifier	FPR	Precision	Recall	F1-measure	MCC	AUC	Acc	Kappa
Bagging	0.018±0.021	0.453±0.441	0.220±0.253	0.266±0.267	0.295±0.251	0.920±0.042	0.508±0.058	0.378±0.077
Random Forest	0.015±0.010	0.167±0.194	0.527±0.237	0.586±0.208	0.580±0.203	0.949±0.048	0.570±0.034	0.539±0.044
AdaBoost M1	0.408±0.207	0.650±0.335	0.307±0.177	0.364±0.133	0.398±0.142	0.905±0.071	0.507±0.027	0.374±0.026
LogitBoost	0.025±0.009	0.167±0.096	0.347±0.169	0.392±0.147	0.286±0.137	0.906±0.072	0.570±0.022	0.492±0.028
XGBoost	0.028±0.019	0.644±0.221	0.513±0.254	0.514±0.175	0.514±0.144	0.878±0.126	0.616±0.030	0.519±0.037

5.4 Effect of Feature Selection

This section presents the performance of feature selection using wrapper (Table 10) and filter (Table 11) feature selection methods in terms of classification measures.

For the wrapper feature selection, an average of 12.6 features were selected, while 3.6 features were selected using the filter approach, indicating that the filter approach identified several redundant features. For the wrapper method, Unemployment, NationalUnemployment, AccualFinancingSurplus/TR, NetOperatingBalance/OR, and NWC/TE were identified most relevant, while for the filter method, NationalUnemployment and OperBalance/OR were selected in all five runs of experiments.

For the wrapper approach, Random forest stands out as the most accurate. It also excels in Cohen's Kappa score, reflecting a robust agreement between its predictions and the actual classifications. SVM, BN and AdaBoost M1 are characterized by an exceptionally low FPR, which underlines their ability to minimize false positive predictions. LogitBoost is notable for its superior Precision, demonstrating the high reliability of its positive classifications. Furthermore, Random forest shows impressive Recall and also leads in the F1 measure, highlighting an optimal balance between precision and recall. Both LogitBoost and Random forest are recognized for their high MCC values, and Bagging and BN are recognized for their high AUC values, demonstrating their performance across different rating classes. For the filter feature selection method, Random forest and XGBoost performed best. Overall, however, no significant improvement was achieved using the feature selection methods for the used classifiers (using the Wilcoxon signed-rank test at $P<0.05$), with few exceptions, such as BN and XGBoost for AUC. This suggests that feature selection is not effective for this problem, indicating that the variables provided by the rating agency are relevant to credit rating and are considered in the evaluation process. Therefore, feature selection was not considered in further experiments.

Table 10: Results of classifiers in combination with wrapper feature selection

Classifier	FPR	Precision	Recall	F1-measure	MCC	AUC	Acc	Kappa
DT	0.041±0.037	0.360±0.234	0.353±0.257	0.330±0.218	0.316±0.193	0.733±0.030	0.517±0.072	0.403±0.084
NN	0.005±0.011	0.300±0.447	0.113±0.176	0.146±0.208	0.162±0.222	0.642±0.222	0.428±0.063	0.261±0.102
SVM	0.00±0.00	0.200±0.477	0.033±0.075	0.057±0.128	0.079±0.177	0.603±0.152	0.309±0.035	0.056±0.074
NB	0.071±0.083	0.336±0.391	0.353±0.282	0.306±0.301	0.287±0.309	0.809±0.064	0.321±0.073	0.191±0.077
BN	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.867±0.089	0.476±0.059	0.342±0.078
JRIP	0.008±0.011	0.200±0.274	0.113±0.176	0.139±0.202	0.135±0.193	0.679±0.237	0.456±0.100	0.295±0.144
FURIA	0.033±0.039	0.300±0.263	0.347±0.265	0.337±0.208	0.343±0.200	0.661±0.122	0.440±0.063	0.416±0.062
Random Forest	0.025±0.024	0.424±0.276	0.393±0.266	0.432±0.239	0.430±0.236	0.858±0.147	0.555±0.0400	0.446±0.050
Bagging	0.025±0.025	0.491±0.379	0.287±0.223	0.317±0.206	0.329±0.193	0.880±0.094	0.515±0.057	0.395±0.068
AdaBoost M1	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.636±0.092	0.281±0.014	0.040±0.037
LogitBoost	0.013±0.013	0.610±0.418	0.347±0.206	0.439±0.224	0.452±0.221	0.821±0.114	0.496±0.055	0.368±0.062
XGBoost	0.040±0.045	0.496±0.405	0.387±0.289	0.369±0.269	0.364±0.276	0.845±0.166	0.529±0.063	0.415±0.076

Table 11: Results of classifiers in combination with filter feature selection

Classifier	FPR	Precision	Recall	F1-measure	MCC	AUC	Acc	Kappa
DT	0.025±0.013	0.249±0.276	0.133±0.138	0.156±0.182	0.176±0.166	0.722±0.091	0.511±0.040	0.389±0.050
NN	0.005±0.011	0.067±0.149	0.033±0.075	0.044±0.099	0.039±0.088	0.677±0.081	0.365±0.061	0.177±0.094
SVM	0.000±0.000	0.200±0.4777	0.033±0.075	0.057±0.128	0.079±0.177	0.513±0.033	0.311±0.039	0.044±0.046
NB	0.076±0.099	0.135±0.132	0.267±0.303	0.167±0.167	0.140±0.147	0.778±0.048	0.330±00.108	0.179±0.118
BN	0.010±0.022	0.040±0.089	0.033±0.075	0.036±0.081	0.025±0.057	0.900±0.045	0.487±0.078	0.359±0.094
JRIP	0.015±0.034	0.200±0.447	0.033±0.075	0.057±0.128	0.094±0.172	0.794±0.113	0.461±0.055	0.300±0.080
FURIA	0.023±0.044	0.190±0.325	0.167±0.236	0.170±0.264	0.155±0.256	0.590±0.140	0.442±0.059	0.297±0.067
Random forest	0.030±0.015	0.501±0.104	0.413±0.187	0.435±0.146	0.412±0.136	0.881±0.086	0.541±0.046	0.434±0.061
Bagging	0.031±0.023	0.217±0.217	0.142±0.189	0.163±0.160	0.160±0.134	0.850±0.117	0.494±0.028	0.366±0.040
AdaBoost M1	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.663±0.047	0.291±0.023	0.046±0.058
LogitBoost	0.010±0.014	0.267±0.435	0.067±0.091	0.102±0.141	0.127±0.171	0.861±0.087	0.482±0.046	0.351±0.056
XGBoost	0.035±0.031	0.444±0.171	0.320±0.177	0.346±0.142	0.325±0.147	0.894±0.070	0.548±0.052	0.439±0.067

5.5 Effect of Class Balancing

Table 12 and Table 13 below demonstrate the experiments while using SMOTE and random oversampling for class balancing, respectively.

Table 12 shows the results for SMOTE. The Random Forest algorithm demonstrates an improved Recall and F1 Measure compared to previous results, indicating better performance after using SMOTE for class balancing. However, Precision remains low. The DT algorithm shows improved TPR, but Precision is still moderate. The use of SMOTE may have helped in handling imbalanced classes. For NN, Recall has improved, and Precision is higher compared to previous results, indicating better performance with class balancing. SVM shows significant improvement in Recall, Precision, and F1 measure after using SMOTE. It performs well in handling imbalanced classes. AdaBoost M1 has also shown improvements in Recall and Precision, indicating better performance with SMOTE. LogitBoost has improved Precision, but other metrics show only moderate improvement. Naïve Bayes has led to improvements in Recall, Precision, and F1 measure through the use of SMOTE. Bayes Net shows moderate improvement in Recall and Precision. Bagging has improved Recall, but Precision remains moderate. FURIA also shows a significant improvement in Recall, Precision, and F1 measures. Overall, the use of SMOTE has generally improved the performance of the classifiers.

As presented in Table 13, Random forest shows improvement in terms of Recall, F1 measure, and MCC. DT shows improved Recall and moderate Precision. The resampling technique seems to have a positive impact. The performance has been positively affected by oversampling for NN. However, SVM shows zero values for various metrics, indicating that resampling might not have been effective for SVM in this context. Resampling may also not have been effective for AdaBoost in this context. LogitBoost has improved Precision, but other metrics show only moderate improvement. Naïve Bayes has shown significant improvement in F1 measure. Bayes Net has shown moderate improvement in Recall and Precision. JRIP shows improvement in Recall, Precision, and F1 measure, and FURIA also shows improvement in terms of these metrics. Resampling techniques have generally been found to have a positive effect on classifier performance, particularly in addressing imbalances and improving sensitivity.

Table 12: Results of classification methods enhanced with SMOTE

Classifier	FPR	Precision	Recall	F1-measure	MCC	AUC	Acc	Kappa
DT	0.083±0.022	0.279±0.137	0.487±0.3347	0.348±0.203	0.309±0.239	0.703±0.176	0.447±0.083	0.416±0.104
NN	0.068±0.019	0.347±0.112	0.487±0.152	0.404±0.127	0.359±0.142	0.886±0.057	0.550±0.059	0.445±0.075
SVM	0.093±0.036	0.376±0.074	0.727±0.083	0.496±0.072	0.472±0.079	0.895±0.067	0.487±0.041	0.352±0.050
NB	0.210±0.265	0.860±0.142	0.560±0.296	0.561±0.266	0.641±0.165	0.922±0.058	0.543±0.032	0.451±0.038
BN	0.030±0.021	0.563±0.140	0.527±0.265	0.513±0.188	0.499±0.180	0.919±0.044	0.475±0.020	0.354±0.020
JRIP	0.063±0.050	0.426±0.124	0.520±0.240	0.424±0.133	0.405±0.123	0.774±0.129	0.464±0.042	0.319±0.051
FURIA	0.074±0.052	0.347±0.150	0.827±0.199	0.612±0.171	0.610±0.172	0.915±0.077	0.503±0.047	0.416±0.056
Random Forest	0.035±0.006	0.167±0.124	0.327±0.316	0.561±0.218	0.540±0.227	0.933±0.028	0.570±0.028	0.499±0.037
Bagging	0.103±0.160	0.398±0.271	0.340±0.259	0.356±0.245	0.345±0.211	0.679±0.094	0.412±0.045	0.245±0.071
AdaBoost M1	0.055±0.024	0.454±0.143	0.587±0.277	0.483±0.131	0.463±0.154	0.896±0.050	0.527±0.061	0.412±0.083
LogitBoost	0.033±0.021	0.167±0.253	0.493±0.253	0.480±0.168	0.286±0.148	0.902±0.083	0.570±0.025	0.470±0.036
XGBoost	0.028±0.011	0.567±0.091	0.527±0.237	0.529±0.186	0.508±0.177	0.855±0.143	0.600±0.039	0.501±0.047

Table 13: Results of classification enhanced with random oversampling

Classifier	FPR	Precision	Recall	F1-measure	MCC	AUC	Acc	Kappa
DT	0.059±0.028	0.404±0.134	0.513±0.096	0.446±0.114	0.406±0.121	0.827±0.098	0.496±0.054	0.416±0.071
NN	0.046±0.021	0.458±0.218	0.487±0.192	0.468±0.196	0.429±0.214	0.903±0.067	0.557±0.036	0.453±0.043
SVM	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.693±0.037	0.302±0.025	0.096±0.049
NB	0.005±0.007	0.910±0.124	0.520±0.267	0.618±0.236	0.646±0.188	0.934±0.049	0.569±0.049	0.416±0.062
BN	0.020±0.007	0.587±0.187	0.487±0.304	0.514±0.268	0.499±0.263	0.905±0.059	0.508±0.058	0.392±0.072
JRIP	0.043±0.026	0.475±0.317	0.353±0.195	0.350±0.135	0.336±0.146	0.738±0.104	0.391±0.058	0.237±0.058
FURIA	0.033±0.019	0.470±0.227	0.387±0.205	0.420±0.216	0.387±0.229	0.736±0.133	0.520±0.044	0.416±0.058
Random Forest	0.033±0.026	0.167±0.396	0.400±0.384	0.387±0.326	0.377±0.340	0.684±0.192	0.570±0.023	0.298±0.031
Bagging	0.015±0.010	0.220±0.303	0.240±0.251	0.275±0.273	0.271±0.251	0.876±0.106	0.513±0.034	0.385±0.050
AdaBoost M1	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.659±0.140	0.335±0.045	0.098±0.055
LogitBoost	0.040±0.027	0.167±0.291	0.347±0.121	0.360±0.072	0.286±0.074	0.843±0.122	0.570±0.022	0.425±0.030
XGBoost	0.043±0.023	0.530±0.085	0.660±0.284	0.560±0.191	0.546±0.182	0.839±0.171	0.581±0.040	0.486±0.051

5.6 Results of Ordinal Classification

In the further set of experiments, the ordinal nature of the target class was considered by using ordinal class classifiers. That is, the individual and ensemble classifiers were used in combination with a simple ordinal classification approach (Frank and Hall, 2001).

The results in Table 14 show a mixed performance of the classifiers in predicting sub-sovereign credit ratings, with no single classifier excelling in all metrics. The DT shows modest performance across the metrics, with an F1 measure and MCC indicating reasonable balance and correlation respectively. The NB classifier outperforms others in terms of Precision, Recall, F1 measure, MCC and Kappa, suggesting a strong ability to correctly identify positive instances and maintain a good balance between Precision and Recall. Its high AUC and Acc also confirm its robustness in distinguishing between classes and its overall reliability. However, the standard deviations indicate some variability in performance, which is crucial for the stability and consistency of the predictions.

Table 14: Results of ordinal class classifiers

Classifier	FPR	Precision	Recall	F1-measure	MCC	AUC	Acc	Kappa
DT	0.045±0.014	0.427±0.075	0.487±0.192	0.443±0.132	0.410±0.128	0.782±0.151	0.518±0.013	0.404±0.011
NN	0.040±0.018	0.505±0.142	0.513±0.096	0.499±0.075	0.467±0.086	0.876±0.089	0.581±0.037	0.481±0.044
SVM	0.000±0.000	0.400±0.548	0.067±0.091	0.114±0.157	0.157±0.217	0.533±0.045	0.496±0.022	0.354±0.022
NB	0.028±0.029	0.667±0.216	0.667±0.312	0.622±0.214	0.622±0.214	0.874±0.111	0.426±0.037	0.416±0.035
BN	0.225±0.434	0.552±0.439	0.240±0.188	0.303±0.202	0.335±0.202	0.844±0.051	0.459±0.047	0.324±0.060
JRIP	0.018±0.025	0.736±0.367	0.254±0.194	0.311±0.112	0.363±0.120	0.602±0.135	0.415±0.091	0.259±0.119
FURIA	0.018±0.019	0.713±0.278	0.380±0.139	0.483±0.120	0.534±0.161	0.735±0.100	0.553±0.034	0.416±0.048
Random Forest	0.020±0.014	0.167±0.207	0.447±0.183	0.499±0.147	0.501±0.109	0.911±0.068	0.570±0.021	0.495±0.032
Bagging	0.018±0.017	0.367±0.415	0.133±0.139	0.182±0.178	0.204±0.180	0.853±0.142	0.504±0.071	0.374±0.093
AdaBoost M1	0.028±0.019	0.480±0.365	0.313±0.251	0.331±0.211	0.451±0.325	0.887±0.059	0.508±0.041	0.381±0.060
LogitBoost	0.025±0.027	0.167±0.354	0.280±0.159	0.329±0.104	0.349±0.109	0.869±0.074	0.570±0.057	0.373±0.073
XGBoost	0.028±0.023	0.606±0.257	0.380±0.139	0.420±0.099	0.420±0.072	0.882±0.097	0.584±0.022	0.482±0.029

On the other hand, the SVM shows a peculiar performance with an FPR of 0.000, indicating no false positives, but significantly lower scores in other metrics, particularly in Recall and F1 measure, suggesting a limited ability to correctly identify positive instances. The Naive Bayes (NB) classifier shows high Precision and Recall, but with considerable variability as indicated by the standard deviations. Ensemble methods such as Random Forest, XGBoost and AdaBoost M1 show high AUC values, indicating their effectiveness in classification tasks, with XGBoost and Random Forest also achieving high Acc and Kappa values, indicating their reliability and agreement with true classifications beyond chance.

Overall, only the DT and FURIA classifiers achieved a significant improvement over their non-ordinal counterparts, suggesting that ensemble methods effectively address the ordinal class problem without the need for ordinal class modifications.

5.7 Misclassification Costs and Different Forecasting Horizons

Table 15 shows the performance of the model in terms of the average results of five training/testing data split. Recall that the data was split into 67% training and 33% testing to avoid overfitting of the machine learning methods. A high AUC value indicates good algorithm performance, even on imbalanced classes. Sub-sovereign credit ratings were also assigned average Misclassification cost, as defined in Table 5. The Random Forest and SVM algorithms were the most effective for both forecasting horizons, resulting in a classification cost almost two times lower than that of Naïve Bayes. As expected, the cost increased as the forecasting horizon extended (for Rating +2). However, the average Misclassification cost was still low enough, below 1, indicating that the average Misclassification was less than one class away. In this experiment, Stacking was also used as a heterogeneous ensemble method, combing all the other classifiers, both single and ensemble. Random forest was still the best, suggesting that its variance reduction strategy is more effective than boosting methods.

Table 15: Performance of classifiers in terms of average cost and AUC

	Rating t + 1		Rating t + 2	
	Misclassification cost	AUC	Misclassification cost	AUC
LR	0.861	**0.866***	0.890	**0.869***
DT	0.971	0.735	0.939	0.725
BN	0.914	**0.876***	**0.850***	**0.895***
SVM	**0.751***	**0.881***	**0.825***	**0.890***
NN	1.045	0.820	1.102	0.838
NB	1.457	0.801	1.354	0.797
Random Forest	**0.747***	**0.886***	**0.835***	**0.870***
AdaBoost M1	1.714	0.714	1.467	0.701
Bagging	0.971	**0.874***	0.906	**0.892***
Stacking	1.718	0.430	4.731	0.436

* Performs significantly better at P<0.05 (Wilcoxon signed-rank test)

5.8 Cost-Sensitive Classification

The MetaCost classifier was used to implement the cost-sensitive classification scheme. Again, Misclassification cost from Table 5 were used to produce the objective function. Table 16 demonstrates that the ensemble learning algorithms, when combined with the MetaCost classifier and Random forest, outperformed the other methods. In particular, MetaCost with Random forest achieved significantly better results than the other single and ensemble classifiers. However, as a single classifier, Random forest outperformed the other ensemble classifiers trained in the MetaCost framework. Therefore, it is confirmed that Random forest is the benchmark method in this domain. In addition, when used in conjunction with MetaCost, it can result in substantial savings in misclassification expenses. This can significantly affect bond investors and the associated interest rates. The effective performance on this imbalanced dataset can be attributed to the SMOTE oversampling technique, which mitigates the risk of overfitting and balances the classes. Overall, the results suggest that this is a cutting-edge approach used in this field.

Table 16: Misclassification cost and area under the curve

Method	Misclassification cost	AUC
MetaCost + AdaBoost M1	1.635±0.213	0.594±0.045
MetaCost + Bagging	0.635±0.082	0.823±0.031
MetaCost + Voting	0.650±0.080	0.838±0.033
MetaCost + LogitBoost	0.643±0.107	0.837±0.018
MetaCost + Decorate	0.584±0.067	**0.856±0.021***
MetaCost + Random Forest	**0.492±0.063***	**0.886±0.021***
MetaCost + SVM	0.843±0.119	0.611±0.022
MetaCost + BN	0.775±0.070	0.792±0.033
MetaCost + NB	0.804±0.087	0.743±0.047
MetaCost + LR	0.641±0.077	0.836±0.029
MetaCost + DT	0.708±0.145	0.733±0.047

* Performs significantly better at $P<0.05$ (Wilcoxon signed-rank test)

5.9 SHAP Explanation Models

The best-performing Random forest model was used to explain the sub-sovereign credit rating model. Global SHAP values were used for model explanations. Figure 6 demonstrates how the model's variables affect sub-sovereign rating classification. A high GDP indicates a strong and robust economy, which is a fundamental factor in credit ratings as it reflects the country or region's ability to generate revenue, repay debt, and maintain fiscal discipline. A larger GDP can often result in higher tax revenues for the government. This revenue generation enables the government to meet its financial obligations, including debt servicing and public services. This contributes to fiscal stability, which is a key consideration for credit rating agencies. Higher levels of unemployment can have a direct impact on an individual's ability to repay debts, such as mortgages, car loans, and credit cards. This, in turn, can affect their ability to repay sub-sovereign debt. A decrease in income can make it challenging to meet financial obligations, resulting in delayed payments or defaults. This can result in a decrease in credit scores, indicating a higher risk to lenders and potentially affecting an individual's access to credit and the interest rates they are offered. Additionally, a higher operating balance, also known as an operating surplus or positive operating balance, can have a positive impact on the credit ratings of sub-sovereign entities. The operating balance is the difference between a government's revenues from its day-to-day activities, such as taxes and fees, and its operating expenses, excluding interest payments on debt. A positive operating balance indicates that the government is generating sufficient revenue to cover its operating expenses, leaving room to allocate funds for debt servicing. Credit rating agencies evaluate an entity's capacity to fulfil its debt obligations. A higher operating balance can increase confidence in the government's ability to service its debt. The impact of an increase in Total External Debt Per Capita (TEPerCapita) on sub-sovereign credit ratings can be complex and depend on various factors. Overall, the results indicate that CRAs consider a range of economic, financial, and structural indicators when assessing a region's creditworthiness. CRAs evaluate a region's creditworthiness by considering a variety of economic and fiscal indicators, including the impact of an increase in the cash surplus to total revenue ratio (CashSurp/TR) and real gross domestic product (GDP). The effect of an increase in CashSurp/TR on sub-sovereign credit ratings can vary depending on several factors, and credit rating agencies analyze a range of indicators when assessing a region's creditworthiness.

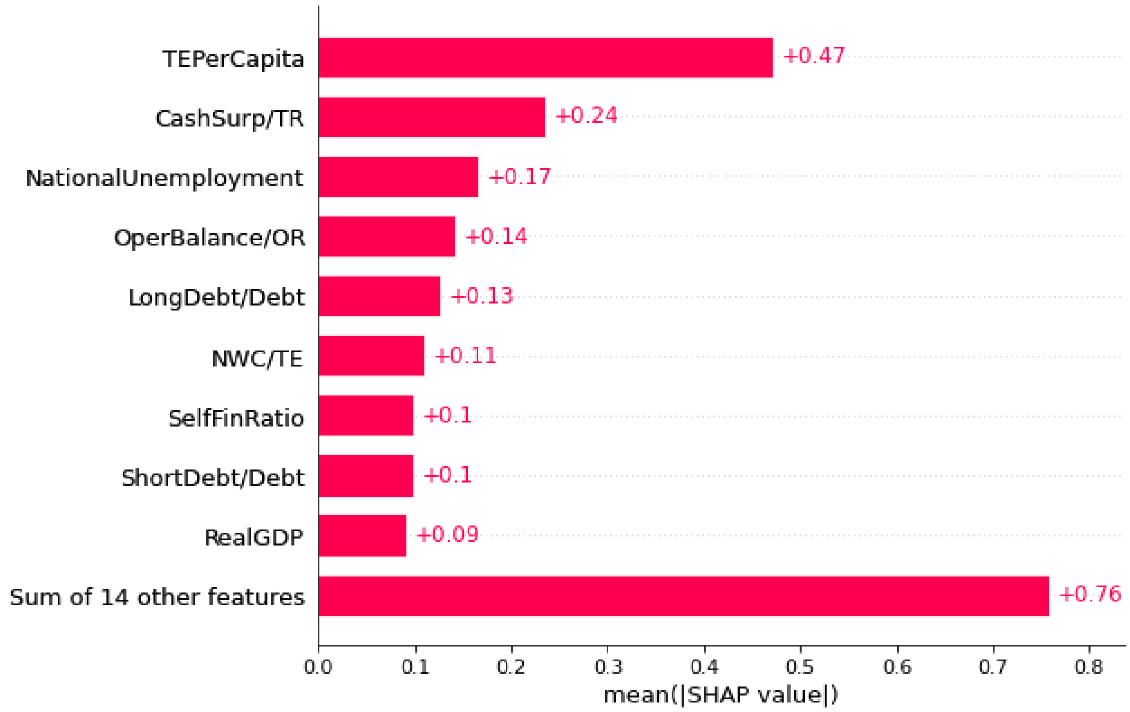


Figure 6: SHAP values for input variables

Figure 7 below demonstrates the impact of the variables for all testing data, also showing direction of the effect. Again, the variables at the top of the graph are the most important, with total expenditure per capita 'TEPerCapita' having the largest positive impact on the credit rating classification. This suggests that higher expenditure per capita (i.e., a region's size) is associated with a higher credit rating. Similarly, cash surplus is higher for better rating classes on the output of the model. Conversely, 'NationalUnemplyoment', 'GovTransf/OperRev' and 'TE/GDP' appear to have a negative impact on the credit rating prediction, as their SHAP values increase with the model's output, indicating that higher unemployment, government transfers, and total expenditure relative to GDP is likely to lower the credit rating.

The spread of SHAP values for each variable indicates the variability in the impact of that feature across different observations in the testing set. For example, "CashSurp/TR" has a mixture of red and blue points on both sides of the zero line, indicating that it can either increase or decrease the credit rating prediction, depending on the specific value. The presence of a high number of points for a feature such as "National Unemployment" suggests that the feature has a variable impact on the model's predictions across different observations, while a more concentrated cluster would indicate a more consistent impact.

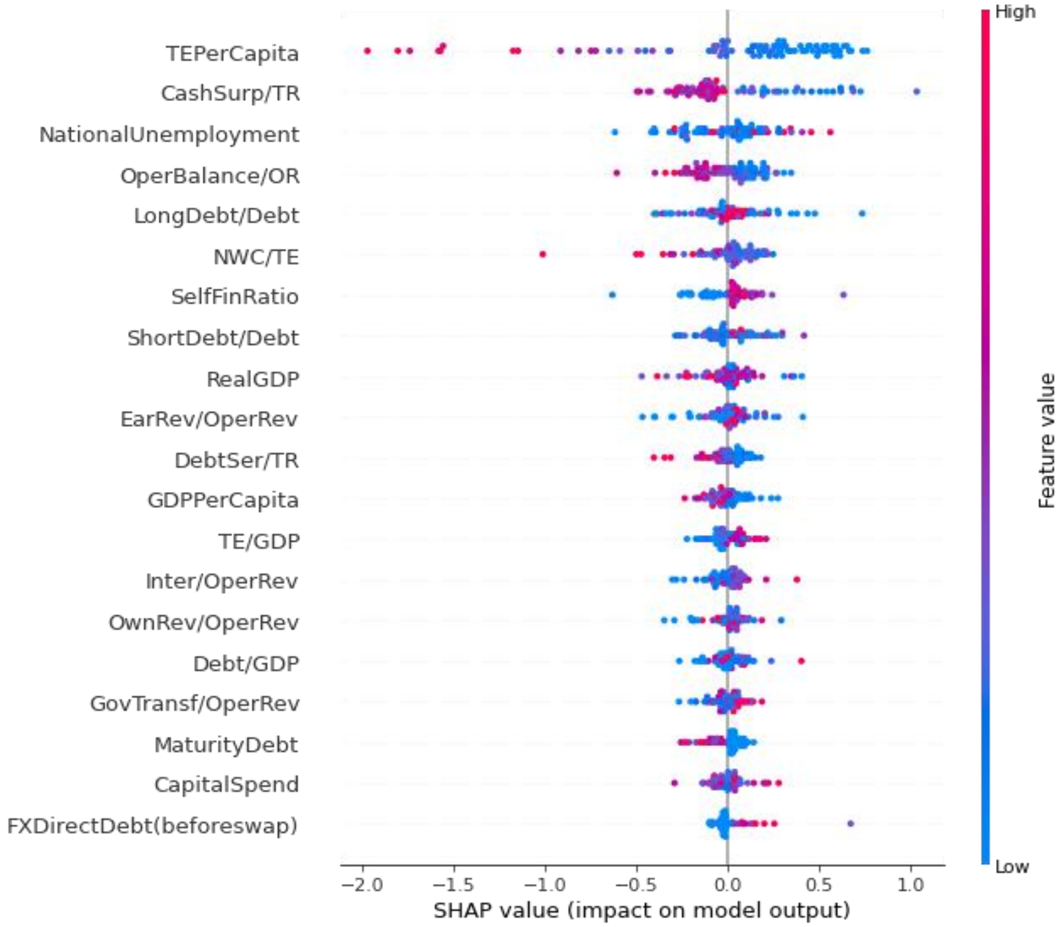


Figure 7: SHAP values (impact on model output)

6 Limitations and Further Research Suggestions

The dissertation thesis was limited to machine learning methods based on supervised learning because all the input and output variables in the datasets were labelled with classes. However, previous studies have also used unlabeled data and employed unsupervised or semi-supervised learning methods (Kennedy, 2013). It is possible that including additional unlabeled data may improve the performance of the proposed models. Collecting additional data is strongly recommended and appears to be a promising approach for future research.

Additionally, this thesis was limited in several ways ranging from the unavailability of more recent data, the choice of dataset, time horizon to the approaches used to develop the models. For instance, this study employed a time series dataset between 2003 and 2016 which might not reflect the current economic or market situation. Besides, the credit rating landscape has changed significantly, with new regulations, industry practices, and economic conditions. It's important to ensure that the data is still relevant and representative of the current credit rating environment. Using more recent data would provide a more accurate reflection of the creditworthiness of regions. The results strongly suggest that the machine learning models based on current available financial and economic data could present accurate classifications of credit ratings. Another limitation of this study was the linear cost matrix. Different distances from default should be considered in the future research to make the cost matrix more realistic. For instance, future researchers could consider the actual cost (interest rates and default costs) of rating misclassification.

Also, the study merely focused on traditional crisp rule-based systems such as Moody's credit rating system to assess the creditworthiness of regions. This limits the study from including alternative credit systems and perspectives that could provide valuable insights. It would be good to consider multiple credit rating systems and approaches in future research to gain a comprehensive understanding. Further, neural networks did not perform well for the two datasets. Hence, future researchers could employ some other models of neural networks, such as MLP with deep learning or ensembles of NNs to improve the poor performance. In a similar vein, ensembles of fuzzy rule-based systems could be a promising direction for future researchers. Also, the proposed model gives an easier idea of the evaluation of the sub-sovereign credit rating for public administration managers, banks, investors, or rating agencies. In future, such models should be developed in a way to make it possible to precisely predict future credit ratings. This can be realized by the combination of feature selection and, consequently, by the classification of the sub-sovereigns into rating classes using different soft computing methods, including those based on fuzzy rules. Moreover, it is imperative to examine several feature selections in order to reveal the relevant / redundant input variables. In future, such models should be developed in a way to make it possible to precisely forecast credit rating. This can be realized by using input variables significant for the credit rating process and, consequently, by the classification of the sub-sovereigns into rating classes. I recommend this for future research.

7 Contributions of the Dissertation Thesis

The aim of the dissertation was to design a new hybrid model by effectively combining various soft computing methods, including their ensembles. Various machine learning methods were compared to resolve a real-world imbalanced multiclass classification task. This determined the use of cost sensitivity and development of an accurate decision support system for credit rating classification of sub-sovereign entities across countries and world regions. The scientific and application contributions of this dissertation thesis are outlined below.

7.1 Scientific Contributions

The scientific contributions of the dissertation thesis include:

- Employing a novel hybrid model based on the effective combination of different soft computing methods, including their ensembles, for forecasting regional financial performance.
- A novel hybrid model designed to determine the cost sensitivity of different machine learning methods and to develop an accurate decision support system that minimizes the cost of rating sub-sovereign entities.
- A novel hybrid model that combines data oversampling with cost-sensitive ensemble classification by employing the SMOTE technique to balance multi-class data effectively to solve the imbalance problem.
- Assigning different misclassification costs in the cost matrix solves the problem of ordered classes. The approach is combined with ensemble classification within the MetaCost framework. This approach demonstrated more accurate predictions in terms of average cost and AUC.
- Investigation of the effect of data preprocessing techniques on the performance of the existing classification methods across different soft-computing methods. Such a comparative study is unique in the existing literature.
- Novel hybrid classification models using ensemble algorithms with different algorithms as base classifiers. Combining multiple base classifiers in a heterogeneous model helps increase the performance and robustness over single model. In contrast to previous ensemble models using DTs as base classifiers, the proposed approach exploits the

advantages of different models in a voting scenario, resolving the problem with high variance in the data.

- Benchmark the proposed hybrid classification models against existing state-of-the-art classification methods. The results demonstrate that the proposed models performed better than the state-of-the-art methods in terms of the prediction criteria.
- For the first time, benchmark datasets were used from multiple regions across the globe. This statement provides robust evidence to support the findings of the dissertation thesis.

7.2 Application Contributions

The dissertation thesis contributes to the field in the following ways:

- Forecasting regional performance using soft computing methods has various applications across different domains. Soft computing techniques, such as fuzzy systems, neural networks and ensemble methods, are particularly useful when dealing with complex, uncertain, and imprecise information. Below are some examples of the applications of forecasting regional performance using soft-computing methods.
- Governments can utilize the proposed hybrid model to forecast regional economic performance, aiding in the formulation of policies for economic development, job creation, and infrastructure planning.
- Investors and financial institutions can use the proposed model to forecast the economic performance of different regions. This information guides investment decisions in stocks, bonds, real estate, and other assets.
- The model can be applied in the financial sector to assess the risk associated with loans and investments in different regions. With this approach, financial institutions can make informed decisions based on objective risk assessments thereby assisting financial institutions to make more informed lending decisions.

Conclusion

Regional financial performance expressed in terms of credit ratings contributes greatly to the efficiency of financial markets. Sub-sovereign credit ratings have become an important source of information for financial market participants and regulators. In this thesis, two comprehensive financial datasets of sub-sovereign entities over the periods of 2003–2007 and 2015 were collected from Moody's rating agency together with the corresponding credit ratings for 2008–2009 and 2016, respectively. The study was designed to determine the misclassification cost of various machine learning methods to consider the ordinal character of rating classes. The main aim was to develop an accurate decision support system that minimizes misclassification cost of credit rating classification for sub-sovereign entities across countries and world regions. I looked at each side of the economic, financial and debt indicators to provide enough inputs to the machine learning models.

The results strongly suggest that the machine learning models based on current available financial and economic data could present accurate classifications of credit ratings. Even though the rating agencies and many other institutional writers' stress the importance of subjective analyses in determining the ratings, it seemed that a small list of input variables largely determine the rating results. This also asserts that the set of variables discovered in this study represent the most relevant information for the credit rating decision. However, in my future research, I will examine several feature selection methods in order to reveal the relevant / redundant input variables.

The results of the first experiment also indicate that the regional financial performance can be accurately predicted at least two years in advance. In the future, therefore, it would be interesting to investigate longer forecasting horizons and update the datasets to cover different economic periods. The results from the second experiments showed that the performance of ensemble learning algorithms can be further improved by using the MetaCost classifier because the ordinal classes are considered. However, the cost matrix was developed in a simplified manner. Different distances from default will be considered in the future research to make the cost matrix more realistic.

The proposed model gives an easier idea of the evaluation of the sub-sovereign credit rating for public administration managers, banks, investors, or rating agencies. In future, such models should be developed in a way to make it possible to precisely predict future credit rating. This can be realized by the combination of feature selection and, consequently, by the classification of the sub-sovereigns into rating classes using different soft computing methods, including those based on fuzzy rules. So far, I only focused on traditional crisp rule-based systems and their fuzzy extensions. Surprisingly, neural networks did not perform well for the two datasets. In an effort to improve this poor performance, I plan to employ some other models of neural networks, such as MLP with deep learning or ensembles of NNs. In a similar vein, ensembles of fuzzy rule-based systems could be a promising direction of the research for my dissertation.

References

[1] Abakah, A.A. 2020. Local religious beliefs and municipal bond market outcomes. *Financial Management*, Vol. 49, No. 2, 447-471.

[2] Adelson, M.H., 2012. *The role of credit ratings in the financial system*. Standard & Poor's.

[3] Aktan, B., Çelik, Ş., Abdulla, Y., and Alshakhoori, N. 2019. The impact of credit ratings on capital structure. *ISRA International Journal of Islamic Finance*, Vol. 11, No. 2, 226-245.

[4] Al-Najjar, D., and Al-Najjar, B. 2014. Developing a multi stage predicting system for corporate credit rating in emerging markets: Jordanian case. *Journal of Enterprise Information Management*, Vol. 27, No. 4, 475-487.

[5] Alsakka, R., and Ap Gwilym, O. 2013. Rating agencies' signals during the European sovereign debt crisis: Market impact and spillovers. *Journal of Economic Behavior & Organization*, Vol. 85, 144-162.

[6] Ammar, S., Duncombe, W., Hou, Y., Jump, B., and Wright, R. 2001. Using fuzzy rule-based systems to evaluate overall financial performance of governments: An enhancement to the bond rating process. *Public Budgeting & Finance*, Vol. 21, No. 4, 91-110.

[7] Amstad, M., and Packer, F. 2015. Sovereign ratings of advanced and emerging economies after the crisis. *BIS Quarterly Review*, Vol. 4, 77-91.

[8] Anand, A., Baesens, B. and Vanpee, R., 2023. Sovereign credit risk modeling using machine learning: a novel approach to sovereign credit risk incorporating private sector and sustainability risks. *Journal of Credit Risk*, Vol. 19, No. 1, 105-154

[9] Antulov-Fantulin, N., Lagravinese, R., and Resce, G. 2021. Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior & Organization*, Vol. 183, 681-699.

[10] Balios, D., Thomadakis, S. and Tsipouri, L., 2016. Credit rating model development: An ordered analysis based on accounting data. *Research in International Business and Finance*, Vol. 38, 122-136,

[11] Basu, R., Chen, X. and Naughton, J. 2022. *Municipal bond market effects of credit rating dissemination*. Available at https://hdl.handle.net/10125/104063

[12] Beck R, Ferrucci G, Hantzsche A., and Rau-Goehring M. 2017. Determinants of sub-sovereign bond yield spreads - The role of fiscal fundamentals and federal bailout expectations. *Journal of International Money and Finance*, Vol. 79, 72-98.

[13] Bennell, J.A., Crabbe, D., Thomas, S., and Ap Gwilym, O. 2006. Modelling sovereign credit ratings: Neural networks versus ordered probit. *Expert Systems with Applications*, Vol. 30, No. 3, 415-425.

[14] Bhatia, M.A.V. 2002. *Sovereign credit ratings methodology: An evaluation*, No. 2-170. International Monetary Fund.

[15] Borensztein, M. E., Valenzuela, P., and Cowan, K. 2007. *Sovereign ceilings "lite"? The impact of sovereign ratings on corporate ratings in emerging market economies*, No. 7-75. International Monetary Fund.

[16] Breiman, L. 1996. Bagging predictors. *Machine Learning*, Vol. 24, No. 2, 123-140.

[17] Breiman, L. 2001. Random forests. *Machine Learning*, Vol. 45, No. 1, 5-32.

[18] Boughaci, D. and Alkhawaldeh, A.A.S. 2018. Three local search-based methods for feature selection in credit scoring. *Vietnam Journal of Computer Science*, Vol. 5, 107-121.

[19] Bretscher, L., Kung, H., and Sahay, A. 2022. *State fiscal policy and municipal credit risk premia*. Available at SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4151915

[20] Buendía-Carrillo, D., Lara-Rubio, J., Navarro-Galera, A., and Gómez-Miranda, M. E. 2020. The impact of population size on the risk of local government default. *International Tax and Public Finance*, Vol. 27, 1264-1286.

[21] Camanho, N., Deb, P. and Liu, Z., 2022. Credit rating and competition. *International Journal of Finance & Economics*, Vol. 27, No. 3, 2873-2897.

[22] Cantor, R., and Packer, F. 1994. The credit rating industry. *Quarterly Review*, Vol. 19, 1-26.

[23] Cao, Z., Kim, J.B., Zhang, E.X., and Zhang, R. 2019. *Managerial risk tolerance and corporate credit ratings*. Available at http://hdl.handle.net/10125/64878

[24] Carmeli, A. 2003. Introduction: Fiscal and financial crises of local governments. *International Journal of Public Administration*, Vol. 26, No. 13, 1423-1430.

[25] Cash, D., 2018. Can credit rating agencies play a greater role in corporate governance disclosure?. *Corporate Governance: The International Journal of Business in Society*, Vol. 18, No. 5, 954-964.

[26] Chakraborty, S., Sharma, R.K. and Tewari, P. 2017. Application of soft computing techniques over hard computing techniques: A survey. *Int. J. Indestructible Math. Comput*., Vol. 1, No. 1, 8-17.

[27] Chalak, K. and Kim, D. 2022. *Estimating the feedback among credit rating agencies and its impact on the municipal bond market*. Available at SSRN 4176998. https://doi.org/10.2139/ssrn.4176998

[28] Chandrashekar, G. and Sahin, F. 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, Vol. 40, No. 1, 16-28.

[29] Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, Vol. 16, 321-357.

[30] Cheikh, N.B., Hmiden, O.B., Zaied, Y.B., and Boubaker, S. 2021. Do sovereign credit ratings matter for corporate credit ratings?. *Annals of Operations Research*, Vol. 297, 77-114.

[31] Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[32] Cheng, C. S. 2020. *A machine learning based bond rating prediction system facilitating investment decision making*. Doctoral dissertation, The George Washington University.

[33] Chen, G., and Chen, J. 2015. A novel wrapper method for feature selection and its applications. *Neurocomputing*, Vol. 159, 219-226.

[34] Chen, Y. 2021. Research on cost-sensitive classification methods for imbalanced data. In *2021 International Conference on Artificial Intelligence, Big Data and Algorithms* (CAIBDA) (pp. 224-228). IEEE.

[35] Chi, W., Wu, S.J. and Zheng, Z. 2020. Determinants and consequences of voluntary corporate social responsibility disclosure: Evidence from private firms. *The British Accounting Review*, Vol. 52, No. 6, p.100939.

[36] Cho, H., Han, S.Y., Oh, S., and Sunwoo, H.Y. 2020. Optimistic credit rating and its influence on corporate decisions: evidence from Korea. *Asia-Pacific Journal of Accounting & Economics*, Vol. 27, No. 5, 612-629.

[37] Chun, A.L., Namvar, E., Ye, X. and Yu, F. 2019. Modeling municipal yields with (and without) bond insurance. *Management Science*, Vol. 65, No. 8, 3694-3713.

[38] Clark, B.Y. 2015. Evaluating the validity and reliability of the financial condition index for local governments. *Public Budgeting & Finance*, Vol. 35, No. 2, 66-88.

[39] Cohen, S., Doumpos, M., Neofytou, E., and Zopounidis, C. 2012. Assessing financial distress where bankruptcy is not an option: An alternative approach for local municipalities. *European Journal of Operational Research*, Vol. 21, No. 1, 270-279.

[40] Cohen, S., Costanzo, A., and Manes-Rossi, F. 2017. Auditors and early signals of financial distress in local governments. *Managerial Auditing Journal*, Vol. 32, No. 3, 234-250.

[41] Crosby, A., and Robbins, D. 2013. Mission impossible: Monitoring municipal fiscal sustainability and stress in Michigan. *Journal of Public Budgeting, Accounting & Financial Management*, Vol. 25, No. 3, 425-445.

[42] de Oliveira, D.S.P., and Montes, G.C. 2023. Forecasting sovereign risk perception of Brazilian bonds: an evaluation of machine learning prediction accuracy. *International Journal of Emerging Markets*, Vol. 18, No. 10, 3414-3436.

[43] Domingos, P. 1999. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155-164.

[44] Dougherty, J., Kohavi, R., and Sahami, M. 1995. Supervised and unsupervised discretization of continuous features. In *Machine Learning Proceedings* 1995 (pp. 194-202). Morgan Kaufmann.

[45] Doumpos, M., and Figueira, J.R. 2019. A multicriteria outranking approach for modeling corporate credit ratings: An application of the Electre Tri-nC method. *Omega*, Vol. 82, 166-180.

[46] Eijffinger, S.C. 2012. Rating agencies: Role and influence of their sovereign credit risk assessment in the eurozone. JCMS: *Journal of Common Market Studies*, Vol. 50, No. 6, 912-921.

[47] Ekmekcioglu, M., Kaya, T., and Tokmakcioglu, K. 2022. Predicting sovereign credit ratings using machine learning algorithms. In *Global Joint Conference on Industrial Engineering and Its Application Areas* (pp. 52-61). Cham: Springer Nature Switzerland

[48] Feng, B., Xue, W., Xue, B., and Liu, Z. 2020, December. Every corporation owns its image: Corporate credit ratings via convolutional neural networks. In *2020 IEEE 6th International Conference on Computer and Communications (ICCC)* (pp. 1578-1583). IEEE.

[49] Fernández, A., Garcia, S., Herrera, F., and Chawla, N.V. 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, Vol. 61, 863-905.

[50] Finlay, S. 2011. Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research, Vol. 210, No. 2, 368-378.

[51] Flandreau, M., and Flores, J.H. 2009. Bonds and brands: foundations of sovereign debt markets, 1820-1830. *The Journal of Economic History*, Vol. 69, No. 3, 646-684.

[52] Frank, E., and Hall, M. 2001. A simple approach to ordinal classification. In *Machine Learning: ECML 2001: 12th European Conference on Machine Learning* Freiburg, Germany, September 5–7, 2001 Proceedings 12 (pp. 145-156). Springer Berlin Heidelberg.

[53] Freund, Y., and Schapire, R.E. 1996. Experiments with a new boosting algorithm. In *13th Internationl Conference on Machine Learning (ICML)*, pp. 148-156.

[54] Friedman, J., Hastie, T., and Tibshirani, R. 2000. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, Vol. 28, No. 2, 337-407.

[55] Galil, K., Hauptman, A., and Rosenboim, R.L, 2023. Prediction of corporate credit ratings with machine learning: Simple interpretative models. *Finance Research Letters*, Vol. 58, p.104648.

[56] Gaillard, N. 2006. *Determinants of Moody's and S&P's subsovereign credit ratings*. World Bank, Fondation Nationale Des Sciences Politiques, France.

[57] Gaillard, N. 2009. The determinants of Moody's sub-sovereign ratings. *International Research Journal of Finance and Economics*, Vol. 31, No. 1, 194-209.

[58] Greer, R.A. 2016. Local government risk assessment: The effect of government type on credit rating decisions in Texas. *Public Budgeting & Finance*, Vol. 36, No. 2, 70-90.

[59] Gennaioli, N., Martin, A., and Rossi, S. 2018. Banks, government bonds, and default: What do the data say?. *Journal of Monetary Economics*, Vol. 98, 98-113.

[60] Gillette, J.R., Samuels, D., and Zhou, F.S. 2020. The effect of credit ratings on disclosure: Evidence from the recalibration of Moody's municipal ratings. *Journal of Accounting Research*, Vol. 58, No. 3, 693-739.

[61] Gmehling, P., and La Mura, P. 2016. A Bayesian inference model for the credit rating scale. *The Journal of Risk Finance*, Vol. 17, No. 4, 390-404.

[62] Golbayani, P., Florescu, I., and Chatterjee, R. 2020. A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *The North American Journal of Economics and Finance*, Vol. 54, p.101251.

[63] Goldmann, L. 2022. *New approaches to corporate credit rating prediction*. The University of Edinburgh, Edinburgh.

[64] Gorina, E., Maher, C., and Joffe, M. 2018. Local fiscal distress: measurement and prediction. *Public Budgeting & Finance*, Vol. 38, No. 1, 72-94.

[65] Gousario, F., and Dharmastuti, C. F. 2015. Regional financial performance and human development index based on study in 20 counties/cities of level I region. *The Winners*, Vol. 16, No. 2, 152-165.

[66] Gregori, W. D., and Marattin, L. 2019. Determinants of fiscal distress in Italian municipalities. *Empirical Economics*, Vol. 56, 1269-1281.

[67] Gu, Q., Lin, W., Sheng, Y. and Wang, H. 2022. Corporate Private Information and Credit Ratings: Evidence from Corporate Geographic Distance. *Emerging Markets Finance and Trade*, 58(11), pp.3215-3232.

[68] Gupta, R. 2023. Financial determinants of corporate credit ratings: An Indian evidence. *International Journal of Finance & Economics*, Vol. 28, No. 2, 1622-1637.

[69] Hajek, P. 2011. Municipal credit rating modelling by neural networks. *Decision Support Systems*, Vol. 51, No. 1, 108-118.

[70] Hajek, P. 2018. Predicting corporate investment/non-investment grade by using interval-valued fuzzy rule-based systems-a cross-region analysis. *Applied Soft Computing*, Vol. 62, 73-85.

[71] Hajek, P., and Michalak, K. 2013. Feature selection in corporate credit rating prediction. *Knowledge-Based Systems*, 51, 72-84.

[72] Hájek, P., and Olej, V. 2011. Credit rating modelling by kernel-based approaches with supervised and semi-supervised learning. *Neural Computing and Applications*, Vol. 20, 761-773.

[73] Hajek, P., Olej, V., and Prochazka, O. 2017. Predicting corporate credit ratings using content analysis of annual reports-a naive bayesian network approach. In *Enterprise Applications, Markets and Services in the Finance Industry: 8th International Workshop*, FinanceCom 2016, Frankfurt, Germany, December 8, 2016, Revised Papers 8 (pp. 47-61). Springer International Publishing.

[74] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R. and Friedman, J. 2009. Random forests. The elements of statistical learning: Data mining, inference, and prediction, pp.587-604.

[75] Herrmann, J. 2020. *Managing for results: How higher municipal credit ratings increase data driven management in cities*. Johns Hopkins University, Baltimore.

[76] Hmiden, O.B., and Cheikh, N.B. 2016. Debt-threshold effect in sovereign credit ratings: New evidence from nonlinear panel smooth transition models. *Finance Research Letters*, Vol. 19, 273-278.

[77] Huang, H., Svec, J., and Wu, E. 2021. The game changer: Regulatory reform and multiple credit ratings. Journal of Banking & Finance, 133, 106279.

[78] Hu, X., Huang, H., Pan, Z. and Shi, J. 2019. Information asymmetry and credit rating: A quasi-natural experiment from China. *Journal of Banking & Finance*, Vol. 106, 132-152.

[79] Hühn, J., and Hüllermeier, E. 2009. FURIA: an algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, 19, 293-319.

[80] Huo, Y., and Gong, B. 2022. Credit rating in dynamic response to the nature of firms and the business model of rating agencies: evidence from the Chinese bond market. *Heliyon*, Vol. 8, No. 12, p.e11884.

[81] Ibrahim, D. 2016. An overview of soft computing. *Procedia Computer Science*, Vol. 102, 34-38.

[82] Ioannou, S., Wójcik, D., and Pažitka, V. 2021. Financial centre bias in sub-sovereign credit ratings. Journal of I*nternational Financial Markets, Institutions and Money*, Vol. 70, p.101261.

[83] Jadhav, S., He, H., and Jenkins, K. 2018. Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing*, Vol. 69, 541-553.

[84] Jia, B., Wu, J., Du, J., Ji, Y., and Zhu, L. 2021. A prediction model for the secure issuance scale of Chinese local government bonds. *Kybernetes*, Vol. 50, No. 5, 1125-1143.

[85] Kaur, J., Vij, M., and Chauhan, A.K. 2023. Signals influencing corporate credit ratings-a systematic literature review. *Decision*, Vol. 50, 91-114.

[86] Kaur, J., Vij, M., and Chauhan, A.K. 2023. Intellectual structure, themes and disciplines of credit rating determination by rating agencies-a bibliometric analysis. *American Journal of Finance and Accounting*, Vol. 7, No. 2, 125-144.

[87] Kennedy, K. 2013. *Credit scoring using machine learning*. Technology University Dublin, Dublin.

[88] Khashman, A. 2009. A neural network model for credit risk evaluation. *International Journal of Neural Systems*, Vol. 19, No. 04, 285-294.

[89] Kiesel, F., and Lücke, F. 2019. ESG in credit ratings and the impact on financial markets. *Financial Markets, Institutions & Instruments*, Vol. 28, No. 3, 263-290.

[90] Kim, S., and Park, S. 2016. Credit rating inflation during the 2000s: lessons from the US State Governments. *International Journal of Economics and Financial Issues*, Vol. 6, No. 1, 13-19.

[91] Kloha, P., Weissert, C. S., and Kleine, R. (2005). Developing and testing a composite model to predict local fiscal distress. *Public Administration Review*, Vol. 65, No. 3, 313-323.

[92] Koerniadi, H. 2023. Corporate risk-taking after changes in credit rating. *International Journal of Managerial Finance*, Vol. 19, No. 1, 48-62.

[93] Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI*, Vol. 14, No. 2, 1137-1145.

[94] Koutanaei, F.N., Sajedi, H., and Khanbabaei, M. 2015. A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. Journal of Retailing and Consumer Services, Vol. 27, 11-23.

[95] Kumar Roy, P., Shaw, K., and Ishizaka, A. (2023). Developing an integrated fuzzy credit rating system for SMEs using fuzzy-BWM and fuzzy-TOPSIS-Sort-C. *Annals of Operations Research*, Vol. 325, 1197-1229.

[96] Kumar, P. R., and Ravi, V. 2007. Bankruptcy prediction in banks and firms via statistical and intelligent techniques - A review. *European Journal of Operational Research*, Vol. 180, No. 1, 1-28.

[97] Lara-Rubio, J., Rayo-Cantón, S., Navarro-Galera, A., and Buendia-Carrillo, D. 2017. Analysing credit risk in large local governments: an empirical study in Spain. *Local Government Studies*, Vol. 43, No. 2, 194-217.

[98] Lee, W.C., Shen, J., Cheong, T.S., and Wojewodzki, M. 2021. Detecting conflicts of interest in credit rating changes: a distribution dynamics approach. *Financial Innovation*, Vol. 7. No. 1, 1-23.

[99] Leng, J., Valli, C., and Armstrong, L. 2010. A wrapper-based feature selection for analysis of large data sets.

[100] Leiser, S., and Mills, S. 2019. Local government fiscal health: Comparing self-assessments to conventional measures. Public Budgeting & Finance, Vol. 39, No. 3, 75-96.

[101] Liang, D., Tsai, C. F., Lu, H. Y. R., and Chang, L. S. 2020. Combining corporate governance indicators with stacking ensembles for financial distress prediction. *Journal of Business Research*, Vol. 120, 137-146.

[102] Li, C., Pervaiz, K., Asif Khan, M., Ur Rehman, F., and Oláh, J. 2019. On the asymmetries of sovereign credit rating announcements and financial market development in the European region. *Sustainability*, Vol. 11, No. 23, p.6636.

[103] Li, J.P., Mirza, N., Rahat, B., and Xiong, D. 2020. Machine learning and credit ratings prediction in the age of fourth industrial revolution. *Technological Forecasting and Social Change*, Vol. 161, p.120309.

[104] Lin, C.M., Chen, C.C.S., Yang, S.Y., and Wang, W.R. 2020. The effects of corporate governance on credit ratings: The role of corporate social responsibility. *Emerging Markets Finance and Trade*, Vol. 56, No. 5, 1093-1112.

[105] Liu, G. 2012. Municipal bond insurance premium, credit rating, and underlying credit risk. *Public Budgeting & Finance*, Vol. 32, No. 1, 128-156.

[106] Liu, H., Zhou, M., and Liu, Q. 2019. An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*, Vol. 6, No. 3, 703-715.

[107] Liu, L., and Tan, K.S. 2009. *Subnational credit ratings: A comparative review*. The World Bank.

[108] Lu, H., Liyan, H., and Hongwei, Z. 2013. Credit scoring model hybridizing artificial intelligence with logistic regression. *Journal of Networks*, Vol. 8, No. 1, 253-261.

[109] Lukac, J., Teplicka, K., Culkova, K., and Hrehova, D. 2021. Evaluation of the financial performance of the municipalities in Slovakia in the context of multidimensional statistics. *Journal of Risk and Financial Management*, Vol. 14, No. 12, p.570.

[110] Lundberg, S.M., and Lee, S.I. 2017. *Consistent feature attribution for tree ensembles*. arXiv preprint arXiv:1706.06060.

[111] Lundberg, S.M., and Lee, S.I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, Vol. 30, 1-10.

[112] Luo, C. 2022. A comparison analysis for credit scoring using bagging ensembles. *Expert Systems*, Vol. 39, No. 2, p.e12297.

[113] Iyer, K.V. and Ravi Kumar, V.V. 2022. Using blockchain technology in credit rating industry to promote an innovative bond-pays model. *International Journal of Innovation Science*, Vol. 14, No. 3/4, 537-555.

[114] Madhav, V.V. 2021. An empirical study on Impact of Credit Rating on Credit Risk of Banks: A Literature Review. *Indian Journal of Finance and Banking*, Vol. 7, No. 1, 31-37.

[115] Maher, C.S., Deller, S.C., Stallmann, J.I., and Park, S. 2016. The impact of tax and expenditure limits on municipal credit ratings. *The American Review of Public Administration*, Vol. 46, No. 5, 592-613.

[116] Maher, C. S., Oh, J. W., and Liao, W. J. 2020. Assessing fiscal distress in small county governments. Journal of Public Budgeting, *Accounting & Financial Management*, Vol. 32, No. 4, 691-711.

[117] Maran, L., Funnell, W., and Castellini, M. 2019. Accounting and the enactment of power: municipal reform by Peter Leopold 1774-1775. *Accounting, Auditing & Accountability Journal*, Vol. 32, No. 4, 1146-1174.

[118] Masungini, A.W., Maseko, G.J., and Robbetze, N. 2023. Evaluating the implementation of a municipal credit management policy. *Journal of Local Government Research and Innovation*, Vol. 4, p.101.

[119] Matias Gama, A.P. and Susana Amaral Geraldes, H. 2012. Credit risk assessment and the impact of the New Basel Capital Accord on small and medium-sized enterprises: An empirical analysis. *Management Research Review*, Vol. 35, No. 8, 727-749.

[120] Michie, D., Spiegelhalter, D.J., and Taylor, C.C., 1994. Machine learning. *Neural and Statistical Classification*, Vol. 13, 1-298.

[121] Mohapatra, S., Nose, M., and Ratha, D. 2018. Determinants of the distance between sovereign credit ratings and sub-sovereign bond ratings: Evidence from emerging markets and developing economies. *Applied Economics*, Vol. 50, No. 9, 934-956.

[122] Muñoz-Izquierdo, N., Segovia-Vargas, M.J., Camacho-Miñano, M.D.M., and Pérez-Pérez, Y. 2022. Machine learning in corporate credit rating assessment using the expanded audit report. *Machine Learning*, Vol. 111, No. 11, 4183-4215.

[123] Nakashima, H.H., Mantovani, D., and Machado Junior, C. 2022. Users' trust in black-box machine learning algorithms. *Revista de Gestão*, in press. Doi: 10.1108/REGE-06-2022-0100

[124] Navarro-Galera, A., Buendia-Carrillo, D., Lara-Rubio, J., and Rayo-Canton, S. (2017). Do political factors affect the risk of local government default? Recent Evidence from Spain. *Lex Localis-Journal of Local Self-Government*, Vol. 15, No. 1, 43-66.

[125] Nilsson, M., and Shan, Q. 2018. *Credit risk analysis with machine learning techniques in peer-to-peer lending market*. Stockholm Business School, Stockholm.

[126] Ntsalaze, Z., Boako, G., and Alagidede, P. 2017. The impact of sovereign credit ratings on corporate credit ratings in South Africa. *African Journal of Economic and Management Studies*, Vol. 8, No. 2, 126-146.

[127] Nwogugu, M.I. and Nwogugu, M.I. 2021. Economic psychology, geopolitical risk and the unconstitutionality of private-sector credit rating agencies, ratings opinions and government bailouts/bail-ins. *Geopolitical Risk, Sustainability and "Cross-Border Spillovers" in Emerging Markets*, Volume II: Constitutional Political Economy, Pandemics-Governance And Labor-Oriented Bail-outs/Bail-ins, pp.145-247.

[128] Olowookere, A.T., and Adewale, S.O. 2020. A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach. *Scientific African*, Vol. 8, 2468-2276,

[129] Omolaye, P. O., Mom, J. M., and Igwue, G. A. 2017. A holistic review of soft computing techniques. *Applied and Computational Mathematics*, Vol. 6, No. 2, 93-110.

[130] Omstedt, M. 2020. Reading risk: The practices, limits and politics of municipal bond rating. *Environment and Planning A: Economy and Space*, Vol. 52. No. 3, 611-631.

[131] Opitz, D., and Maclin, R., 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, Vol. 11, 169-198.

[132] Overes, B.H., and van der Wel, M. 2023. Modelling sovereign credit ratings: Evaluating the accuracy and driving factors using machine learning techniques. *Computational Economics*, Vol. 61, No. 3, 1273-1303.

[133] Pamuk, M. and Schumann, M. 2023. Opening a New Era with Machine Learning in Financial Services? Forecasting Corporate Credit Ratings Based on Annual Financial Statements. *International Journal of Financial Studies*, 11(3), p.96.

[134] Pan, W. T. 2012. A new fruit fly optimization algorithm: taking the financial distress model as an example. *Knowledge-Based Systems*, Vol. 26, 69-74.

[135] Panetta, F., Correa, R., Davies, M., Di Cesare, A., Marques, J.M., Nadal de Simone, F., and Wieland, M. 2011. *The impact of sovereign credit risk on bank funding*. MPRA Paper, available at https://mpra.ub.uni-muenchen.de/32581/1/ MPRA _paper_32581.pdf

[136] Patel, M.C., Panchal, M., and Bhavsar, H.P. 2013. Decorate ensemble of artificial neural networks with high diversity for classification. *International Journal of Computer Science and Mobile Computing*, Vol. 2, No. 5, 134-138.

[137] Paula, D.A.V.D., Artes, R., Ayres, F., and Minardi, A.M.A.F. 2019. Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques. *RAUSP Management Journal*, Vol. 54, 321-336.

[138] Petropoulos, A., Chatzis, S.P., and Xanthopoulos, S. 2016. A novel corporate credit rating system based on Student'st hidden Markov models. *Expert Systems with Applications*, Vol. 53, 87-105.

[139] Petropoulos, A., Siakoulis, V., Stavroulakis, E. and Klamargias, A. 2019. A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. *IFC Bulletins Chapters*, 49.

[140] Ren, G., Hong, T. and Park, Y. 2015. Multi-Class SVM+ MTL for the prediction of corporate credit rating with structured data. *Asia Pacific Journal of Information Systems*, Vol. 25, No. 3, 579-596.

[141] Reusens, P., and Croux, C. 2017. Sovereign credit rating determinants: A comparison before and after the European debt crisis. *Journal of Banking & Finance*, Vol. 77, 108-121.

[142] Roy, P.K., and Shaw, K. 2023. An integrated fuzzy credit rating model using fuzzy-BWM and new fuzzy-TOPSIS-Sort-C. *Complex & Intelligent Systems*, Vol. 9, No. 4, 3581-3600.

[143] Sajjad, F., and Zakaria, M. 2018. Credit rating as a mechanism for capital structure optimization: Empirical evidence from panel data analysis. *International Journal of Financial Studies*, Vol. 6, No. 1, p.13.

[144] Sarkar, B.K. 2016. A case study on partitioning data for classification. *International Journal of Information and Decision Sciences*, Vol. 8, No. 1, 73-91.

[145] Sharma, C., Singh, A., and Yadav, R. 2023. Impact of competition in credit rating industry: Evidence from India. *SAGE Open*, Vol. 13, No. 1, p.21582440221135107.,

[146] Shengdong, M., Yunjie, L., and Jijian, G. 2023. Entrepreneurial borrowing overdue prediction based on stacking model transfer learning. *Management Decision*, in press. Doi: /10.1108/MD-10-2022-1462

[147] Štrumbelj, E. and Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, Vol. 41, 647-665.

[148] Sun, J., Lang, J., Fujita, H., and Li, H. 2018. Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, Vol. 425, 76-91.

[149] Sun, J., Li, H., Huang, Q. H., and He, K. Y. 2014. Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, Vol. 57, 41-56.

[150] Sun, Y., Chai, N., Dong, Y., and Shi, B. 2022. Assessing and predicting small industrial enterprises' credit ratings: A fuzzy decision-making approach. *International Journal of Forecasting*, Vol. 38, No. 3, 1158-1172.

[151] 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*, Vol. 559, 153-170.

[152] Takawira, O., and Mwamba, J.W.M. 2022. Sovereign credit ratings analysis using the logistic regression model. *Risks*, 10(4), p.70.

[153] Teixeira, J.C., Silva, F.J., Ferreira, M.B., and Vieira, J.A. 2018. Sovereign credit rating determinants under financial crises. *Global Finance Journal*, Vol. 36, 1-13.

[154] Toseafa, E. 2017. Forecasting sub-sovereign credit ratings using machine learning methods. In *Proceedings of the 30th International Business Information Management Association Conference*. International Business Information Management Association-IBIMA.

[155] Toseafa, E. 2018. Using meta learning methods to forecast sub-sovereign credit ratings. *Journal of Eastern Europe Research in Business and Economics*, Vol. 2018, p.870203,

[156] Toseafa, E., and Hajek, P. 2019. Predicting regional credit ratings using ensemble classification with metacost. In *Artificial Intelligence Methods in Intelligent Algorithms: Proceedings of 8th Computer Science On-line Conference 2019*, Vol. 2 8 (pp. 332-342). Springer International Publishing.

[157] Trevino, L., and Thomas, S. 2000. *Systematic differences in the determinants of foreign currency sovereign ratings by rating agency*. University of Southampton, Working Paper No. 00-153.

[158] Tsai, C.F., and Chen, M.L. 2010. Credit rating by hybrid machine learning techniques. *Applied Soft Computing*, Vol. 10, No. 2, 374-380.

[159] Ubarhande, P. and Chandani, A. 2021. Elements of credit rating: a hybrid review and future research Agenda. *Cogent Business & Management*, Vol. 8, No. 1, p.1878977.

[160] Yeh, C. C., Chi, D. J., and Hsu, M. F. 2010. A hybrid approach of DEA, rough set and support vector machines for business failure prediction. *Expert Systems with Applications*, Vol. 37, No. 2, 1535-1541.

[161] Yeh, C. C., Lin, F., and Hsu, C. Y. 2012. A hybrid KMV model, random forests and rough set theory approach for credit rating. *Knowledge-Based Systems*, Vol. 33, 166-172.

[162] Yu, B., Li, C., Mirza, N., and Umar, M. 2022. Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. *Technological Forecasting and Social Change*, Vol. 174, p.121255.

[163] Veeralakshmi, V., and Ramyachitra, D. 2015. Ripple Down Rule learner (RIDOR) classifier for IRIS dataset. *Issues*, Vol. 1, No. 1, 79-85.

[164] Viswanathan, H.H.K.K., Murugesan, P., Rengasamy, S., and Vilvanathan, L. 2021. A modified boosted support vector machine to rate banks. *Benchmarking: An International Journal*, Vol. 28. No. 1, 1-27.

[165] Wah, Y.B., Ibrahim, N., Hamid, H.A., Abdul-Rahman, S., and Fong, S. 2018. Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science & Technology*, Vol. 26, No. 1, 329-340.

[166] Wang, X., Dennis, L., and Tu, Y. S. 2007. Measuring financial condition: A study of US states. *Public Budgeting & Finance*, Vol. 27, No. 2, 1-21.

[167] Wang, M., and Ku, H. 2021. Utilizing historical data for corporate credit rating assessment. *Expert Systems with Applications*, Vol. 165, p.113925.

[168] Wang, Y., Wang, S., and Lai, K. K. 2005. A new fuzzy support vector machine to evaluate credit risk. *IEEE Transactions on Fuzzy Systems*, Vol. 13, No. 6, 820-831.

[169] Weng, C.H., and Huang, C.K. 2021. A hybrid machine learning model for credit approval. *Applied Artificial Intelligence*, Vol. 35, No. 15, 1439-1465.

[170] Wong, B.K., Lai, V.S., and Lam, J. 2000. A bibliography of neural network business applications research: 1994–1998. *Computers & Operations Research*, Vol. 27, No. 11-12, 1045-1076.

[171] Wu, H.C., Hu, Y.H. and Huang, Y.H. 2014. Two-stage credit rating prediction using machine learning techniques. *Kybernetes*, Vol. 43. No. 7, 1098-1113,

[172] Wu, J., Zhang, Z., and Zhou, S. X. 2022. Credit rating prediction through supply chains: a machine learning approach. *Production and Operations Management*, Vol. 31, No. 4, 1613-1629.

[173] Yinger, J. 2010. Municipal Bond Ratings and Citizens' Rights. *American law and economics review*, 12(1), pp.1-38.

[174] Yu, B., Li, C., Mirza, N., and Umar, M. 2022. Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. *Technological Forecasting and Social Change*, Vol. 174, p.121255.

[175] Yu, X., and Fang, J. 2022. Tax credit rating and corporate innovation decisions. *China Journal of Accounting Research*, Vol. 15, No. 1, p.100222.

[176] Zadeh, L. A. 1994. Soft computing and fuzzy logic. *IEEE Software*, Vol. 11, No. 6, 48-56.

[177] Zafra-Gomez, J. L., Lopez-Hernandez, A. M., and Hernández-Bastida, A. 2009. Developing a model to measure financial condition in local government: Evaluating service quality and minimizing the effects of the socioeconomic environment: An application to Spanish municipalities. T*he American Review of Public Administration*, Vol. 39, No. 4, 425-449.

[178] Zhang, R., Zhang, Z., Wang, D., and Du, M. 2022. Financial distress prediction with a novel diversity-considered GA-MLP ensemble algorithm. *Neural Processing Letters*, Vol. 54, No. 2, 1175-1194.

[179] Zhang, R., Chen, B., and Sakdanuphab, R. 2024. An improved neural manufacturing corporate credit rating model based on LSTM. *International Journal of Intelligent Systems and Applications in Engineering*, Vol. 12, No. 2, 338-351.

[180] Zhang, X., Zhu, S., He, J., Yang, B. and Guan, X. 2019. Credit rating based real-time energy trading in microgrids. *Applied Energy*, Vol. 236, 985-996.

[181] Zhuang, K.Z. 2023. *The correlation and effectiveness between two variables that are considered in sustainable investing: ESG ratings and credit ratings*. University of Tennessee, Knoxville.

[182] Zhu, Y., Xie, C., Wang, G.J., and Yan, X.G. 2017. Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications*, Vol. 28, No. 1, 41-50.

[183] Zukanović, M., Milošević, P., Poledica, A., and Vučičević, A. 2022. An approach to corporate credit rating prediction using computational intelligence-based methods. In *International Symposium SymOrg* (pp. 80-95). Cham: Springer International Publishing.

Publications of the student

Journal papers

[1] Jirava, P., and Toseafa, E., 2017. An illustrative case study of the integration of enterprise resource planning system. *Journal of Enterprise Resource Planning Studies*, Vol. 2017, Article 176215. Indexed: EBSCO.

[2] Toseafa, E. 2018. Using meta learning methods to forecast sub-sovereign credit Ratings. Journal of Eastern Europe Research in Business and Economics, Vol. 2018, Article 870203. Indexed: EBSCO.

[3] Toseafa, H., Bata, R., and Toseafa, E. 2018. Incidence of occupational health hazards and safety culture at Tema oil refinery (Tor) in Ghana: Exploring the symbiotic relationship. *British Journal of Environmental Sciences*, Vol. 6, No. 4, 58-74. Indexed: Google Scholar.

[4] Toseafa, E., and Hajek, P., 2019. Predicting regional credit ratings using ensemble classification with metacost. *Advances in Intelligent Systems and Computing*, Vol. 985, 332-342. Indexed: Scopus. SJR: 0.215, Q3.

[5] Odei, S. A., Odei, M. A., and Toseafa, E., 2022. Determinants of technological and non-technological innovations: Evidence from Ghana' manufacturing and service sectors. *Journal of African Business*, Vol. 24, No. 3, 467-490. doi: 10.1080/15228916.2022.2113209. Indexed: WoS (AIS: 0.448), Scopus, SJR: 0.611, Q1.

[6] Appiah, M.K., Toseafa, E., Sam, A., Danso, F., and Nsowah, A., 2023. Implications of sustainability initiatives on African Continental Free Trade adoption by firms under environmental uncertainty. *International Journal of Development Issues*, Vol. 22, No. 2, 255-277.

Conference papers

[7] Jirava, P., and Toseafa, E., 2016. Integration of enterprise resource planning system in the multinational corporation case study. In *Proceedings of the 28th International Business Information Management Association Conference* (pp. 3847-3854). Indexed: WoS.

[8] Toseafa, E., 2017. Forecasting sub-sovereign credit ratings using machine learning methods. In *Proceedings of the 30th International Business Information Management Association Conference* (pp. 1271-1279). Indexed: WoS

APPENDICES

Appendix A

Parameter estimates (cauchit)

Threshold/Variables	Estimate	Std. Error	Wald	*P*-value
[Class_Aaa = 1]	15604.00	2412.00	41871.00	0.000***
[Class_Aa = 2]	8870.00	1600.00	30743.00	0.000***
[Class_A = 3]	5264.00	1243.00	17938.00	0.000***
[Class_Baa = 4]	3072.00	1169.00	6907.00	0.009***
[Class_Ba = 5]	-0.524	1064.00	0.243	0.622
[Class_B = 6]	-6245.00	1894.00	10876.00	0.001***
[Class_Caa = 7]	-50529.00	42002.00	1447.00	0.229
GDPPerCapita	0.004	0.003	1204.00	0.272
RealGDP	0.271	0.101	7147.00	0.008**
Unemployment	0.064	0.047	1854.00	0.173
NationalUnemployment	-0.173	0.055	10012.00	0.002***
DebtGDP	-0.075	0.022	11791.00	0.001***
FXDirectDebtbeforeswap	-0.075	0.014	28650.00	0.000***
ShortDebtDebt	-0.066	0.016	16676.00	0.000***
LongDebtDebt	-0.025	0.006	19264.00	0.000***
MaturityDebt	0.204	0.048	18335.00	0.000***
OwnRevOperRev	0.039	0.009	17288.00	0.000***
GovTransfOperRev	0.045	0.010	21096.00	0.000***
EarRevOperRev	-0.001	0.009	0.001	0.978
InterOperRev	-0.819	0.166	24263.00	0.000***
DebtSerTR	0.202	0.039	27427.00	0.000***
AccualFinancingSurplusTR	-0.052	0.044	1406.00	0.236
CashSurpTR	-0.017	0.037	0.201	0.654
TEPerCapita	0.001	0.000	47292.00	0.000***
TEGDP	-0.098	0.032	9212.00	0.002***
OperBalanceOR	0.143	0.033	19281.00	0.000***
NetOperatingBalanceOR	-0.008	0.014	0.306	0.580
SelfFinRatio	-0.255	0.271	0.888	0.346
CapitalSpend	-0.043	0.021	4242.00	0.039*
NWCTE	0.107	0.017	42146.00	0.000***

*Note: statistically significant at * P < 0.05, ** P < 0.01, *** P < 0.001*

Appendix B

Parameter Estimates (logit)

Threshold/ Variables	Estimate	Std. Error	Wald	*P*-value
[Class_Aaa = 1]	4.941	1.010	23.940	0.000***
[Class_Aa = 2]	2.069	0.889	5.418	0.020***
[Class_A = 3]	0.431	0.867	0.247	0.619
[Class_Baa = 4]	-0.965	0.868	1.238	0.266
[Class_Ba = 5]	-3.159	0.893	12.513	0.000***
[Class_B = 6]	-5.605	1.005	31.110	0.000***
[Class_Caa = 7]	-7.773	1.422	29.860	0.000***
GDPPerCapita	-0.001	0.002	0.027	0.870
RealGDP	0.024	0.082	0.087	0.769
Unemployment	0.034	0.041	0.683	0.408
NationalUnemployment	-0.157	0.044	12.750	0.000***
DebtGDP	-0.005	0.017	0.073	0.786
FXDirectDebtbeforeswap	-0.035	0.010	13.334	0.000***
ShortDebtDebt	-0.025	0.012	4.707	0.030***
LongDebtDebt	-0.014	0.004	11.027	0.001***
MaturityDebt	0.157	0.033	23.204	0.000***
OwnRevOperRev	0.023	0.007	10.917	0.001***
GovTransfOperRev	0.004	0.007	0.295	0.587
EarRevOperRev	0.003	0.008	0.118	0.732
InterOperRev	-0.281	0.115	6.034	0.014**
DebtSerTR	0.133	0.028	21.995	0.000***
AccualFinancingSurplusTR	-0.011	0.036	0.094	0.760
CashSurpTR	0.056	0.032	3.077	0.079
TEPerCapita	-0.001	6.36E-02	14.436	0.000***
TEGDP	-0.046	0.027	2.825	0.093
OperBalanceOR	0.031	0.024	1.653	0.199
NetOperatingBalanceOR	0.027	0.013	3.910	0.048***
SelfFinRatio	-0.628	0.250	6.326	0.012**
CapitalSpend	-0.019	0.018	1.117	0.291
NWCTE	0.040	0.008	23.681	0.000***

*Note: statistically significant at * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$*

Appendix C

Parameters of estimates (Probit)

Threshold/Variable	Estimate	Std. Error	Wald	*P*-value
[Class_Aaa = 1]	-2.058	0.532	14.944	**0.000*** **
[Class_Aa = 2]	-0.597	0.486	1.507	0.220
[Class_A = 3]	0.279	0.481	0.337	0.562
[Class_Baa = 4]	1.052	0.484	4.718	**0.030*** **
[Class_Ba = 5]	2.304	0.498	21.413	**0.000*** **
[Class_B = 6]	3.497	0.536	42.634	**0.000*** **
[Class_Caa = 7]	4.276	0.613	48.617	**0.000*** **
GDPPerCapita	-0.001	0.001	0.420	0.517
RealGDP	0.014	0.047	0.087	0.768
Unemployment	0.023	0.023	1.033	0.310
NationalUnemployment	-0.098	0.025	15.356	**0.000*** **
DebtGDP	-0.001	0.009	0.002	0.966
FXDirectDebtbeforeswap	-0.020	0.005	13.206	**0.000*** **
ShortDebtDebt	-0.012	0.007	3.626	**0.057*** **
LongDebtDebt	-0.007	0.002	9.045	**0.003*** **
MaturityDebt	0.089	0.018	23.886	**0.000*** **
OwnRevOperRev	0.010	0.004	6.597	**0.010*** **
GovTransfOperRev	-0.001	0.004	0.071	0.790
EarRevOperRev	-0.002	0.005	0.145	0.704
InterOperRev	-0.123	0.065	3.632	**0.057*** **
DebtSerTR	0.060	0.016	14.634	**0.000*** **
AccualFinancingSurplusTR	-0.008	0.020	0.175	0.675
CashSurpTR	0.034	0.018	3.556	**0.059*** **
TEPerCapita	-0.001	3.23E-02	13.294	**0.000*** **
TEGDP	-0.028	0.015	3.312	0.069
OperBalanceOR	0.011	0.014	0.679	0.410
NetOperatingBalanceOR	0.016	0.008	4.325	**0.038*** **
SelfFinRatio	-0.366	0.141	6.712	**0.010*** **
CapitalSpend	-0.012	0.010	1.405	0.236
NWCTE	-0.021	0.004	22.144	**0.000*** **

*Note: statistically significant at * P < 0.05, ** P < 0.01, *** P < 0.001*

Appendix D

Experimental results of single classifiers (mean and standard deviation).

RF	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.500	0.667	0.694	0.988	0.663	0.578
2	0.025	0.600	0.500	0.545	0.517	0.907	0.659	0.575
3	0.013	0.800	0.800	0.800	0.787	0.984	0.619	0.518
4	0.025	0.667	0.667	0.667	0.642	0.979	0.640	0.550
5	0.013	0.500	0.167	0.250	0.261	0.888	0.581	0.474
Mean	0.015	0.167	0.527	0.586	0.580	0.949	0.570	0.539
St.Dev.	0.010	0.194	0.237	0.208	0.203	0.048	0.034	0.044

DT	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.100	0.385	0.833	0.526	0.522	0.860	0.500	0.390
2	0.063	0.345	0.500	0.429	0.383	0.712	0.529	0.416
3	0.051	0.500	0.800	0.615	0.604	0.885	0.488	0.369
4	0.088	0.364	0.667	0.471	0.442	0.879	0.570	0.465
5	0.025	0.000	0.000	0.000	0.042	0.481	0.442	0.299
Mean	0.065	0.319	0.560	0.408	0.399	0.763	0.506	0.416
St.Dev.	0.030	0.188	0.339	0.239	0.216	0.173	0.048	0.061

NN	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.333	0.167	0.222	0.197	0.933	0.535	0.425
2	0.051	0.333	0.333	0.333	0.283	0.825	0.541	0.441
3	0.051	0.200	0.200	0.200	0.149	0.929	0.571	0.467
4	0.038	0.500	0.500	0.500	0.463	0.940	0.535	0.422
5	0.000	1.000	0.333	0.500	0.563	0.848	0.605	0.518
Mean	0.033	0.473	0.307	0.351	0.331	0.895	0.557	0.455
St.Dev.	0.021	0.313	0.132	0.145	0.177	0.054	0.031	0.040

SVM	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.167	0.286	0.396	0.969	0.512	0.375
2	0.000	0.000	0.000	0.000	0.000	0.835	0.529	0.406
3	0.000	1.000	0.200	0.333	0.436	0.949	0.464	0.314
4	0.000	0.000	0.000	0.000	0.000	0.950	0.523	0.390
5	0.038	0.250	0.167	0.200	0.156	0.821	0.500	0.359
Mean	0.008	0.450	0.107	0.164	0.198	0.905	0.506	0.369
St.Dev.	0.017	0.512	0.098	0.157	0.210	0.071	0.026	0.035

Adaboost M1	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.167	0.286	0.396	0.969	0.521	0.375
2	0.000	0.000	0.000	0.000	0.000	0.835	0.529	0.406
3	0.000	1.000	0.200	0.333	0.436	0.949	0.464	0.314
4	0.000	0.000	0.000	0.000	0.000	0.950	0.523	0.390
5	0.038	0.250	0.167	0.200	0.156	0.821	0.500	0.359
Mean	0.008	0.450	0.107	0.164	0.198	0.905	0.507	0.369
St.Dev.	0.017	0.512	0.098	0.157	0.210	0.071	0.027	0.035

Logitboost	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.038	0.500	0.500	0.500	0.463	0.963	0.605	0.516
2	0.025	0.333	0.167	0.222	0.196	0.833	0.624	0.527
3	0.025	0.500	0.400	0.444	0.416	0.954	0.583	0.480
4	0.025	0.600	0.500	0.545	0.517	0.960	0.581	0.476
5	0.013	0.500	0.167	0.250	0.261	0.821	0.570	0.461
Mean	0.025	0.167	0.347	0.392	0.286	0.906	0.570	0.492
St.Dev.	0.009	0.096	0.169	0.147	0.137	0.072	0.022	0.028

Naïve Bayes	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.500	0.667	0.694	0.908	0.558	0.480
2	0.013	0.800	0.667	0.727	0.712	0.939	0.565	0.479
3	0.025	0.667	0.800	0.727	0.712	0.982	0.512	0.410
4	0.013	0.833	0.833	0.833	0.821	0.942	0.535	0.432
5	0.000	1.000	0.167	0.286	0.396	0.823	0.593	0.505
Mean	0.010	0.860	0.593	0.648	0.667	0.919	0.553	0.505
St.Dev.	0.011	0.142	0.272	0.211	0.160	0.060	0.031	0.039

Bayes Net	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.333	0.500	0.563	0.955	0.547	0.442
2	0.013	0.000	0.000	0.000	0.030	0.897	0.388	0.244
3	0.025	0.714	1.000	0.833	0.834	0.997	0.512	0.400
4	0.013	0.000	0.000	0.000	0.030	0.902	0.384	0.227
5	0.038	0.250	0.167	0.200	0.156	0.508	0.430	0.293
Mean	0.018	0.393	0.300	0.307	0.323	0.852	0.452	0.321
St.Dev.	0.014	0.447	0.415	0.358	0.360	0.197	0.074	0.095

Bagging	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.013	0.667	0.333	0.444	0.445	0.967	0.500	0.370
2	0.051	0.000	0.000	0.000	0.061	0.905	0.435	0.276
3	0.025	0.600	0.600	0.600	0.575	0.939	0.524	0.399
4	0.000	0.000	0.000	0.000	0.000	0.931	0.488	0.357
5	0.000	1.000	0.167	0.286	0.396	0.856	0.593	0.490
Mean	0.018	0.453	0.220	0.266	0.295	0.920	0.508	0.378
St.Dev.	0.021	0.441	0.253	0.267	0.251	0.042	0.058	0.077

JRIP	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.500	0.333	0.400	0.373	0.821	0.477	0.328
2	0.076	0.143	0.167	0.154	0.085	0.643	0.459	0.313
3	0.051	0.429	0.600	0.500	0.470	0.861	0.500	0.365
4	0.075	0.400	0.667	0.500	0.470	0.804	0.430	0.287
5	0.013	0.667	0.333	0.444	0.445	0.722	0.477	0.346
Mean	0.048	0.428	0.420	0.400	0.369	0.770	0.469	0.328
St.Dev.	0.029	0.190	0.208	0.144	0.163	0.087	0.026	0.030

FURIA	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.500	0.667	0.694	0.700	0.465	0.323
2	0.013	0.667	0.333	0.444	0.445	0.714	0.541	0.429
3	0.038	0.500	0.600	0.545	0.516	0.765	0.512	0.388
4	0.013	0.750	0.500	0.600	0.590	0.811	0.430	0.274
5	0.013	0.500	0.167	0.250	0.261	0.546	0.512	0.401
Mean	0.015	0.683	0.420	0.501	0.501	0.707	0.492	0.416
St.Dev.	0.014	0.207	0.171	0.162	0.163	0.100	0.044	0.063

XGBoost	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.038	0.625	0.833	0.714	0.698	0.979	0.581	0.479
2	0.051	0.429	0.500	0.461	0.419	0.823	0.600	0.499
3	0.025	0.500	0.400	0.444	0.416	0.934	0.655	0.564
4	0.025	0.667	0.667	0.667	0.642	0.973	0.640	0.552
5	0.000	1.000	0.167	0.286	0.396	0.683	0.605	0.501
Mean	0.028	0.644	0.513	0.514	0.514	0.878	0.616	0.519
St.Dev.	0.019	0.221	0.254	0.175	0.144	0.126	0.030	0.037

Appendix E

Experimental results of classifiers used in conjunction with wrapper feature selection.

RF + wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.600	0.500	0.545	0.517	0.947	0.581	0.481
2	0.063	0.286	0.333	0.308	0.252	0.841	0.506	0.381
3	0.013	0.800	0.800	0.800	0.787	0.972	0.607	0.508
4	0.025	0.333	0.167	0.222	0.197	0.921	0.535	0.423
5	0.000	0.100	0.167	0.286	0.396	0.610	0.545	0.438
Mean	0.025	0.424	0.393	0.432	0.430	0.858	0.555	0.446
St.Dev.	0.024	0.276	0.266	0.239	0.236	0.147	0.040	0.050

DT+wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.600	0.500	0.545	0.517	0.733	0.593	0.496
2	0.101	0.273	0.500	0.353	0.304	0.690	0.400	0.266
3	0.051	0.429	0.600	0.500	0.470	0.757	0.524	0.411
4	0.013	0.000	0.000	0.000	0.030	0.764	0.547	0.431
5	0.013	0.500	0.167	0.250	0.261	0.721	0.523	0.409
Mean	0.041	0.360	0.353	0.330	0.316	0.733	0.517	0.403
St.Dev.	0.037	0.234	0.257	0.218	0.193	0.030	0.072	0.084

NN+wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	0.000	0.000	0.000	0.000	0.927	0.442	0.294
2	0.000	0.000	0.000	0.000	0.000	0.553	0.435	0.250
3	0.025	0.500	0.400	0.444	0.416	0.828	0.381	0.223
4	0.000	0.000	0.000	0.000	0.000	0.423	0.360	0.128
5	0.000	1.000	0.167	0.286	0.396	0.481	0.523	0.408
Mean	0.005	0.300	0.113	0.146	0.162	0.642	0.428	0.261
St.Dev.	0.011	0.447	0.176	0.208	0.222	0.222	0.063	0.102

SVM +wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	0.000	0.000	0.000	0.000	0.857	0.360	0.181
2	0.000	0.000	0.000	0.000	0.000	0.475	0.282	0.010
3	0.000	0.000	0.000	0.000	0.000	0.600	0.274	0.023
4	0.000	0.000	0.000	0.000	0.000	0.500	0.302	0.000
5	0.000	1.000	0.167	0.286	0.396	0.583	0.326	0.067
Mean	0.000	0.200	0.033	0.057	0.079	0.603	0.309	0.056
St.Dev.	0.000	0.447	0.075	0.128	0.177	0.152	0.035	0.074

Adaboost M1 + wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	0.000	0.000	0.000	0.000	0.750	0.279	0.073
2	0.000	0.000	0.000	0.000	0.000	0.627	0.282	0.025
3	0.000	0.000	0.000	0.000	0.000	0.671	0.262	0.086
4	0.000	0.000	0.000	0.000	0.000	0.638	0.279	0.017
5	0.000	0.000	0.000	0.000	0.000	0.496	0.302	0.000
Mean	0.000	0.000	0.000	0.000	0.000	0.636	0.281	0.040
St.Dev.	0.000	0.000	0.000	0.000	0.000	0.092	0.014	0.037

LogitBoost +wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.667	0.800	0.806	0.921	0.547	0.437
2	0.025	0.333	0.167	0.222	0.196	0.766	0.447	0.308
3	0.025	0.050	0.400	0.444	0.416	0.868	0.440	0.310
4	0.013	0.667	0.333	0.444	0.445	0.902	0.558	0.427
5	0.000	1.000	0.167	0.286	0.396	0.647	0.488	0.360
Mean	0.013	0.610	0.347	0.439	0.452	0.821	0.496	0.368
St.Dev.	0.013	0.418	0.206	0.224	0.221	0.114	0.055	0.062

Naïve Bayes + wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.667	0.800	0.806	0.890	0.384	0.247
2	0.190	0.118	0.333	0.174	0.092	0.781	0.247	0.103
3	0.127	0.231	0.600	0.333	0.310	0.856	0.310	0.192
4	0.013	0.000	0.000	0.000	0.030	0.728	0.256	0.129
5	0.025	0.333	0.167	0.222	0.197	0.788	0.407	0.286
Mean	0.071	0.336	0.353	0.306	0.287	0.809	0.321	0.191
St.Dev.	0.083	0.391	0.282	0.301	0.309	0.064	0.073	0.077

Bayes Net+wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	0.000	0.000	0.000	0.000	0.897	0.512	0.396
2	0.000	0.000	0.000	0.000	0.000	0.911	0.541	0.428
3	0.000	0.000	0.000	0.000	0.000	0.966	0.500	0.366
4	0.000	0.000	0.000	0.000	0.000	0.825	0.407	0.265
5	0.000	0.000	0.000	0.000	0.000	0.735	0.419	0.255
Mean	0.000	0.000	0.000	0.000	0.000	0.867	0.476	0.342
St.Dev.	0.000	0.000	0.000	0.000	0.000	0.089	0.059	0.078

Bagging +wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.013	0.667	0.333	0.444	0.445	0.942	0.547	0.435
2	0.063	0.286	0.333	0.308	0.252	0.729	0.424	0.286
3	0.038	0.500	0.600	0.545	0.516	0.953	0.500	0.376
4	0.013	0.000	0.000	0.000	0.030	0.929	0.535	0.421
5	0.000	1.000	0.167	0.286	0.400	0.846	0.570	0.456
Mean	0.025	0.491	0.287	0.317	0.329	0.880	0.515	0.395
St.Dev.	0.025	0.379	0.223	0.206	0.193	0.094	0.057	0.068

JRIP + wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	0.000	0.000	0.000	0.000	0.600	0.523	0.382
2	0.000	0.000	0.000	0.000	0.000	0.665	0.282	0.041
3	0.025	0.500	0.400	0.444	0.416	0.904	0.464	0.335
4	0.000	0.000	0.000	0.000	0.000	0.896	0.523	0.380
5	0.013	0.500	0.167	0.250	0.261	0.332	0.488	0.335
Mean	0.008	0.200	0.113	0.139	0.135	0.679	0.456	0.295
St.Dev.	0.011	0.274	0.176	0.202	0.193	0.237	0.100	0.144

FURIA +wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.075	0.333	0.500	0.400	0.354	0.780	0.465	0.334
2	0.076	0.400	0.667	0.500	0.470	0.764	0.329	0.190
3	0.013	0.667	0.400	0.500	0.494	0.690	0.488	0.345
4	0.000	0.000	0.000	0.000	0.000	0.500	0.453	0.305
5	0.000	0.100	0.167	0.286	0.396	0.573	0.465	0.317
Mean	0.033	0.300	0.347	0.337	0.343	0.661	0.440	0.416
St.Dev.	0.039	0.263	0.265	0.208	0.200	0.122	0.063	0.062

XGBoost +wrapper	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.013	0.800	0.667	0.727	0.712	0.965	0.593	0.496
2	0.114	0.250	0.500	0.333	0.284	0.768	0.435	0.304
3	0.051	0.429	0.600	0.500	0.470	0.954	0.548	0.434
4	0.025	0.000	0.000	0.000	-0.042	0.952	0.500	0.378
5	0.000	1.000	0.167	0.286	0.396	0.588	0.570	0.462
Mean	0.040	0.496	0.387	0.369	0.364	0.845	0.529	0.415
St.Dev.	0.045	0.405	0.289	0.269	0.276	0.166	0.063	0.076

Appendix F

Experimental results of classifiers combined with SMOTE.

SVM + SMOTE	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.075	0.400	0.667	0.500	0.470	0.952	0.512	0.384
2	0.152	0.250	0.667	0.364	0.337	0.791	0.459	0.326
3	0.076	0.400	0.800	0.533	0.529	0.933	0.429	0.276
4	0.100	0.385	0.833	0.526	0.522	0.933	0.512	0.383
5	0.063	0.444	0.667	0.533	0.503	0.867	0.523	0.392
Mean	0.093	0.376	0.727	0.491	0.472	0.895	0.487	0.352
St.Dev.	0.036	0.074	0.083	0.072	0.079	0.067	0.041	0.050

Adaboost M1 + SMOTE	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.075	0.500	1.000	0.667	0.680	0.952	0.628	0.548
2	0.063	0.286	0.333	0.308	0.252	0.861	0.482	0.353
3	0.063	0.375	0.600	0.462	0.433	0.906	0.536	0.429
4	0.063	0.444	0.667	0.533	0.503	0.931	0.477	0.346
5	0.013	0.667	0.333	0.444	0.445	0.829	0.512	0.386
Mean	0.055	0.454	0.587	0.483	0.463	0.896	0.527	0.412
St.Dev.	0.024	0.143	0.277	0.131	0.154	0.050	0.061	0.083

LogitBoost +SMOTE	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.038	0.500	0.500	0.500	0.463	0.938	0.581	0.488
2	0.051	0.333	0.333	0.333	0.283	0.852	0.612	0.520
3	0.051	0.500	0.800	0.615	0.604	0.972	0.571	0.467
4	0.025	0.667	0.667	0.667	0.642	0.967	0.558	0.447
5	0.000	1.000	0.167	0.286	0.396	0.781	0.547	0.427
Mean	0.033	0.167	0.493	0.480	0.286	0.902	0.570	0.470
St.Dev.	0.021	0.253	0.253	0.168	0.148	0.083	0.025	0.036

Naive Bayes +SMOTE	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.500	1.000	0.333	0.500	0.563	0.929	0.535	0.453
2	0.013	0.800	0.667	0.727	0.712	0.932	0.553	0.466
3	0.025	0.667	0.800	0.727	0.712	0.980	0.512	0.412
4	0.013	0.833	0.833	0.833	0.821	0.946	0.523	0.418
5	0.500	1.000	0.167	0.167	0.396	0.825	0.593	0.505
Mean	0.210	0.860	0.560	0.591	0.641	0.922	0.543	0.451
St.Dev.	0.265	0.142	0.296	0.266	0.165	0.058	0.032	0.038

Bayes Net+SMOTE	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.500	0.333	0.400	0.373	0.921	0.477	0.357
2	0.013	0.800	0.667	0.727	0.712	0.857	0.459	0.336
3	0.063	0.444	0.800	0.571	0.564	0.971	0.452	0.332
4	0.038	0.571	0.667	0.615	0.586	0.948	0.500	0.375
5	0.013	0.500	0.167	0.250	0.261	0.900	0.488	0.371
Mean	0.030	0.563	0.527	0.513	0.499	0.919	0.475	0.354
St.Dev.	0.021	0.140	0.265	0.188	0.180	0.044	0.020	0.020

Bagging+ SMOTE	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.038	0.000	0.000	0.000	0.052	0.644	0.337	0.127
2	0.388	0.571	0.667	0.615	0.586	0.811	0.412	0.259
3	0.038	0.250	0.200	0.222	0.180	0.551	0.440	0.289
4	0.038	0.500	0.500	0.500	0.463	0.699	0.453	0.307
5	0.013	0.667	0.333	0.444	0.445	0.692	0.419	0.241
Mean	0.103	0.398	0.340	0.356	0.345	0.679	0.412	0.245
St.Dev.	0.160	0.271	0.259	0.245	0.221	0.094	0.045	0.071

JRIP + SMOTE	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.088	0.417	0.833	0.556	0.548	0.900	0.442	0.295
2	0.038	0.500	0.500	0.500	0.462	0.851	0.494	0.352
3	0.139	0.214	0.600	0.316	0.293	0.820	0.512	0.389
4	0.038	0.500	0.500	0.500	0.463	0.728	0.407	0.260
5	0.013	0.500	0.167	0.250	0.261	0.573	0.465	0.298
Mean	0.063	0.426	0.520	0.424	0.405	0.774	0.464	0.319
St.Dev.	0.050	0.124	0.240	0.133	0.123	0.129	0.042	0.051

FURIA+ SMOTE	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.113	0.400	1.000	0.571	0.596	0.967	0.477	0.359
2	0.051	0.500	0.667	0.571	0.540	0.809	0.482	0.359
3	0.076	0.400	0.800	0.533	0.529	0.868	0.512	0.395
4	0.130	0.333	0.833	0.476	0.476	0.933	0.465	0.337
5	0.000	0.100	0.833	0.909	0.907	1.000	0.581	0.479
Aver	0.074	0.347	0.827	0.612	0.610	0.915	0.503	0.416
St.Dev.	0.052	0.150	0.119	0.171	0.172	0.077	0.047	0.056

XGBoost +SMOTE	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.667	0.667	0.667	0.642	0.960	0.547	0.437
2	0.038	0.500	0.500	0.500	0.462	0.810	0.612	0.515
3	0.025	0.667	0.800	0.727	0.712	0.922	0.655	0.568
4	0.038	0.500	0.500	0.500	0.463	0.958	0.593	0.494
5	0.013	0.500	0.167	0.250	0.261	0.623	0.593	0.489
Mean	0.028	0.567	0.527	0.529	0.508	0.855	0.600	0.501
St.Dev.	0.011	0.091	0.237	0.186	0.177	0.143	0.039	0.047

Appendix G

Experimental results of classifiers combined with random oversampling (ROS).

SVM + ROS	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.075	0.400	0.667	0.500	0.470	0.952	0.512	0.384
2	0.152	0.250	0.667	0.364	0.337	0.791	0.459	0.326
3	0.076	0.400	0.800	0.533	0.529	0.933	0.429	0.276
4	0.100	0.385	0.833	0.526	0.522	0.933	0.512	0.383
5	0.063	0.444	0.667	0.533	0.503	0.867	0.523	0.392
Mean	0.093	0.376	0.727	0.491	0.472	0.895	0.487	0.352
St.Dev.	0.036	0.074	0.083	0.072	0.079	0.067	0.041	0.050

AdaBoost M1 + ROS	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.075	0.500	1.000	0.667	0.680	0.952	0.628	0.548
2	0.063	0.286	0.333	0.308	0.252	0.861	0.482	0.353
3	0.063	0.375	0.600	0.462	0.433	0.906	0.536	0.429
4	0.063	0.444	0.667	0.533	0.503	0.931	0.477	0.346
5	0.013	0.667	0.333	0.444	0.445	0.829	0.512	0.386
Mean	0.055	0.454	0.587	0.483	0.463	0.896	0.527	0.412
St.Dev.	0.024	0.143	0.277	0.131	0.154	0.050	0.061	0.083

LogitBoost + ROS	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.038	0.500	0.500	0.500	0.463	0.938	0.581	0.488
2	0.051	0.333	0.333	0.333	0.283	0.852	0.612	0.520
3	0.051	0.500	0.800	0.615	0.604	0.972	0.571	0.467
4	0.025	0.667	0.667	0.667	0.642	0.967	0.558	0.447
5	0.000	1.000	0.167	0.286	0.396	0.781	0.547	0.427
Mean	0.033	0.167	0.493	0.480	0.286	0.902	0.570	0.470
St.Dev.	0.021	0.253	0.253	0.168	0.148	0.083	0.025	0.036

Naïve Bayes +ROS	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.500	1.000	0.333	0.500	0.563	0.929	0.535	0.453
2	0.013	0.800	0.667	0.727	0.712	0.932	0.553	0.466
3	0.025	0.667	0.800	0.727	0.712	0.980	0.512	0.412
4	0.013	0.833	0.833	0.833	0.821	0.946	0.523	0.418
5	0.500	1.000	0.167	0.167	0.396	0.825	0.593	0.505
Mean	0.210	0.860	0.560	0.591	0.641	0.922	0.543	0.451
St.Dev.	0.265	0.142	0.296	0.266	0.165	0.058	0.032	0.038

Bayes Net+ROS	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.500	0.333	0.400	0.373	0.921	0.477	0.357
2	0.013	0.800	0.667	0.727	0.712	0.857	0.459	0.336
3	0.063	0.444	0.800	0.571	0.564	0.971	0.452	0.332
4	0.038	0.571	0.667	0.615	0.586	0.948	0.500	0.375
5	0.013	0.500	0.167	0.250	0.261	0.900	0.488	0.371
Mean	0.030	0.563	0.527	0.513	0.499	0.919	0.475	0.354
St.Dev.	0.021	0.140	0.265	0.188	0.180	0.044	0.020	0.020

Bagging+ ROS	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.038	0.000	0.000	0.000	0.052	0.644	0.337	0.127
2	0.388	0.571	0.667	0.615	0.586	0.811	0.412	0.259
3	0.038	0.250	0.200	0.222	0.180	0.551	0.440	0.289
4	0.038	0.500	0.500	0.500	0.463	0.699	0.453	0.307
5	0.013	0.667	0.333	0.444	0.445	0.692	0.419	0.241
Mean	0.103	0.398	0.340	0.356	0.345	0.679	0.412	0.245
St.Dev.	0.160	0.271	0.259	0.245	0.221	0.094	0.045	0.071

JRIP + ROS	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.088	0.417	0.833	0.556	0.548	0.900	0.442	0.295
2	0.038	0.500	0.500	0.500	0.462	0.851	0.494	0.352
3	0.139	0.214	0.600	0.316	0.293	0.820	0.512	0.389
4	0.038	0.500	0.500	0.500	0.463	0.728	0.407	0.260
5	0.013	0.500	0.167	0.250	0.261	0.573	0.465	0.298
Mean	0.063	0.426	0.520	0.424	0.405	0.774	0.464	0.319
St.Dev.	0.050	0.124	0.240	0.133	0.123	0.129	0.042	0.051

FURIA +ROS	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.038	0.500	500	0.500	0.463	0.877	0.547	0.445
2	0.063	0.167	0.167	0.167	0.103	0.648	0.471	0.341
3	0.013	0.750	0.600	0.667	0.653	0.857	0.536	0.415
4	0.025	0.600	0.500	0.545	0.517	0.731	0.570	0.462
5	0.025	0.333	0.167	0.222	0.197	0.568	0.477	0.339
Mean	0.033	0.470	0.387	0.420	0.387	0.736	0.520	0.146
St.Dev.	0.019	0.227	0.205	0.216	0.229	0.133	0.044	0.058

XGBoost+ ROS	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.050	0.556	0.833	0.667	0.652	0.977	0.628	0.548
2	0.076	0.400	0.667	0.500	0.470	0.778	0.565	0.472
3	0.038	0.571	0.800	0.667	0.652	0.901	0.607	0.520
4	0.038	0.625	0.833	0.714	0.698	0.971	0.581	0.481
5	0.013	0.500	0.167	0.250	0.261	0.569	0.523	0.413
Mean	0.043	0.530	0.660	0.560	0.546	0.839	0.581	0.486
St.Dev.	0.023	0.085	0.284	0.191	0.182	0.171	0.040	0.051

Appendix H

Experimental results of Ordinal class (OC) classifiers.

RF + OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.038	0.500	0.500	0.500	0.463	0.963	0.570	0.460
2	0.025	0.667	0.667	0.667	0.641	0.878	0.624	0.530
3	0.025	0.500	0.400	0.444	0.416	0.954	0.607	0.505
4	0.013	0.750	0.500	0.600	0.590	0.952	0.616	0.517
5	0.000	1.000	0.167	0.286	0.396	0.806	0.612	0.461
Mean	0.020	0.167	0.447	0.499	0.501	0.911	0.570	0.495
St.Dev.	0.014	0.207	0.183	0.147	0.109	0.068	0.021	0.032

DT+OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.050	0.429	0.500	0.462	0.419	0.967	0.523	0.412
2	0.038	0.500	0.500	0.500	0.462	0.741	0.518	0.400
3	0.063	0.375	0.600	0.462	0.433	0.778	0.500	0.388
4	0.050	0.500	0.667	0.571	0.541	0.864	0.512	0.404
5	0.025	0.333	0.167	0.222	0.197	0.561	0.535	0.416
Mean	0.045	0.427	0.487	0.443	0.410	0.782	0.518	0.404
St.Dev.	0.014	0.075	0.192	0.132	0.128	0.151	0.013	0.011

NN+OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.038	0.500	0.500	0.500	0.463	0.904	0.581	0.482
2	0.063	0.444	0.667	0.533	0.502	0.793	0.553	0.454
3	0.038	0.400	0.400	0.400	0.362	0.959	0.583	0.481
4	0.050	0.429	0.500	0.462	0.419	0.952	0.547	0.437
5	0.013	0.750	0.500	0.600	0.590	0.771	0.640	0.552
Mean	0.040	0.505	0.513	0.499	0.467	0.876	0.581	0.481
St.Dev.	0.018	0.142	0.096	0.075	0.086	0.089	0.037	0.044

SVM + OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.167	0.286	0.396	0.583	0.465	0.326
2	0.000	0.000	0.000	0.000	0.000	0.500	0.494	0.346
3	0.000	0.000	0.000	0.000	0.000	0.500	0.512	0.376
4	0.000	0.000	0.000	0.000	0.000	0.500	0.488	0.345
5	0.000	1.000	0.167	0.286	0.396	0.583	0.523	0.375
Mean	0.000	0.400	0.067	0.114	0.158	0.533	0.496	0.354
St.Dev.	0.000	0.548	0.091	0.157	0.217	0.045	0.022	0.022

AdaBoost M1 + OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.500	0.333	0.400	0.373	0.948	0.523	0.401
2	0.025	0.000	0.000	0.000	0.043	0.842	0.482	0.350
3	0.038	0.400	0.400	0.400	0.362	0.908	0.524	0.406
4	0.050	0.500	0.667	0.571	0.541	0.927	0.558	0.454
5	0.000	1.000	0.167	0.286	0.936	0.810	0.453	0.296
Mean	0.028	0.480	0.313	0.331	0.451	0.887	0.508	0.381
St.Dev.	0.019	0.356	0.251	0.211	0.325	0.059	0.041	0.060

LogitBoost+OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.063	0.375	0.500	0.429	0.382	0.915	0.477	0.353
2	0.038	0.250	0.167	0.200	0.156	0.889	0.435	0.288
3	0.025	0.500	0.400	0.444	0.416	0.865	0.548	0.438
4	0.000	1.000	0.167	0.285	0.396	0.931	0.570	0.458
5	0.000	1.000	0.167	0.286	0.396	0.744	0.465	0.327
Mean	0.025	0.167	0.280	0.329	0.349	0.869	0.570	0.373
St.Dev.	0.027	0.354	0.159	0.104	0.109	0.074	0.057	0.073

Naive Bayes+OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.667	0.800	0.806	0.865	0.395	0.288
2	0.025	0.667	0.667	0.667	0.641	0.941	0.412	0.306
3	0.076	0.454	1.000	0.625	0.648	0.982	0.393	0.282
4	0.025	0.714	0.833	0.769	0.753	0.888	0.453	0.348
5	0.013	0.500	0.167	0.250	0.261	0.692	0.477	0.358
Mean	0.028	0.667	0.667	0.622	0.622	0.874	0.426	0.416
St.Dev.	0.029	0.216	0.312	0.220	0.214	0.111	0.037	0.035

Bayes Net+OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.050	0.429	0.500	0.462	0.419	0.855	0.488	0.368
2	0.050	0.000	0.000	0.000	0.061	0.793	0.424	0.284
3	1.000	1.000	0.200	0.333	0.436	0.803	0.488	0.362
4	0.000	1.000	0.333	0.500	0.563	0.921	0.500	0.369
5	0.025	0.333	0.167	0.222	0.197	0.850	0.395	0.237
Mean	0.225	0.552	0.240	0.303	0.335	0.844	0.459	0.324
St.Dev.	0.434	0.439	0.188	0.202	0.202	0.051	0.047	0.060

Bagging+ OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.500	0.333	0.400	0.373	0.829	0.535	0.419
2	0.038	0.000	0.000	0.000	0.053	0.616	0.400	0.237
3	0.000	0.000	0.000	0.000	0.000	0.957	0.595	0.489
4	0.025	0.333	0.167	0.222	0.197	0.946	0.488	0.358
5	0.000	1.000	0.167	0.286	0.396	0.916	0.500	0.365
Mean	0.018	0.367	0.133	0.182	0.204	0.853	0.504	0.374
St.Dev.	0.017	0.415	0.139	0.178	0.180	0.142	0.071	0.093

JRIP + OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.167	0.285	0.396	0.599	0.302	0.099
2	0.038	0.250	0.167	0.200	0.156	0.403	0.506	0.383
3	0.051	0.429	0.600	0.500	0.470	0.780	0.440	0.322
4	0.000	1.000	0.167	0.286	0.396	0.643	0.337	0.169
5	0.000	1.000	0.167	0.286	0.396	0.583	0.488	0.322
Mean	0.018	0.736	0.254	0.311	0.363	0.602	0.415	0.259
St.Dev.	0.025	0.367	0.194	0.112	0.120	0.135	0.091	0.119

FURIA+OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.038	0.500	0.500	0.500	0.463	0.721	0.547	0.430
2	0.038	0.400	0.333	0.364	0.322	0.795	0.518	0.390
3	0.013	0.667	0.400	0.500	0.494	0.777	0.583	0.477
4	0.000	1.000	0.500	0.667	0.694	0.815	0.523	0.391
5	0.000	1.000	0.167	0.386	0.696	0.568	0.593	0.495
Mean	0.018	0.713	0.380	0.483	0.534	0.735	0.553	0.416
St.Dev.	0.019	0.278	0.139	0.120	0.161	0.100	0.034	0.048

XGBoost+OC	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.600	0.500	0.545	0.517	0.963	0.558	0.447
2	0.051	0.429	0.500	0.462	0.419	0.732	0.565	0.460
3	0.051	0.333	0.400	0.364	0.321	0.934	0.592	0.496
4	0.013	0.667	0.333	0.444	0.445	0.946	0.612	0.520
5	0.000	1.000	0.167	0.286	0.396	0.835	0.593	0.487
Mean	0.028	0.606	0.380	0.420	0.420	0.882	0.584	0.482
St.Dev.	0.023	0.257	0.139	0.099	0.072	0.097	0.022	0.029

Appendix I

Experimental results of classifiers combined with filter feature selection.

RF +filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.025	0.600	0.500	0.545	0.517	0.820	0.593	0.502
2	0.025	0.500	0.333	0.400	0.373	0.759	0.506	0.389
3	0.051	0.333	0.400	0.364	0.321	0.953	0.488	0.365
4	0.038	0.571	0.667	0.615	0.586	0.932	0.581	0.492
5	0.013	0.500	0.167	0.250	0.261	0.940	0.535	0.420
Mean	0.030	0.501	0.413	0.435	0.412	0.881	0.541	0.434
St.Dev.	0.015	0.104	0.187	0.146	0.136	0.086	0.046	0.061

DT+filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.013	0.667	0.330	0.444	0.445	0.719	0.500	0.386
2	0.038	0.000	0.000	0.000	0.053	0.747	0.447	0.308
3	0.013	0.000	0.000	0.000	0.028	0.809	0.536	0.408
4	0.038	0.250	0.167	0.167	0.156	0.767	0.523	0.402
5	0.025	0.330	0.167	0.167	0.197	0.570	0.547	0.442
Mean	0.025	0.249	0.133	0.156	0.176	0.722	0.511	0.389
St.Dev.	0.013	0.276	0.138	0.182	0.166	0.091	0.040	0.050

NN+filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	0.000	0.000	0.000	0.000	0.767	0.395	0.218
2	0.000	0.000	0.000	0.000	0.000	0.608	0.259	0.011
3	0.000	0.000	0.000	0.000	0.000	0.752	0.392	0.229
4	0.000	0.000	0.000	0.000	0.000	0.669	0.372	0.190
5	0.025	0.333	0.167	0.222	0.197	0.590	0.407	0.235
Mean	0.005	0.067	0.033	0.044	0.039	0.677	0.365	0.177
St.Dev.	0.011	0.149	0.075	0.099	0.088	0.081	0.061	0.094

SVM+filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	0.000	0.000	0.000	0.000	0.494	0.337	0.076
2	0.000	0.000	0.000	0.000	0.000	0.500	0.259	0.039
3	0.000	0.000	0.000	0.000	0.000	0.500	0.298	0.000
4	0.000	0.000	0.000	0.000	0.000	0.500	0.302	0.000
5	0.000	1.000	0.167	0.286	0.396	0.573	0.360	0.103
Mean	0.000	0.200	0.033	0.057	0.079	0.513	0.311	0.044
St.Dev.	0.000	0.447	0.075	0.128	0.177	0.033	0.039	0.046

AdaBoost M1+filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	0.000	0.000	0.000	0.000	0.725	0.291	0.000
2	0.000	0.000	0.000	0.000	0.000	0.600	0.294	0.000
3	0.000	0.000	0.000	0.000	0.000	0.671	0.262	0.086
4	0.000	0.000	0.000	0.000	0.000	0.638	0.280	0.017
5	0.000	0.000	0.000	0.000	0.000	0.679	0.326	0.128
Mean	0.000	0.000	0.000	0.000	0.000	0.663	0.291	0.046
St.Dev.	0.000	0.000	0.000	0.000	0.000	0.047	0.023	0.058

LogitBoost +filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	1.000	0.167	0.286	0.396	0.813	0.477	0.344
2	0.025	0.000	0.000	0.000	0.043	0.732	0.435	0.294
3	0.000	0.000	0.000	0.000	0.000	0.928	0.440	0.300
4	0.000	0.000	0.000	0.000	0.000	0.896	0.535	0.411
5	0.025	0.333	0.167	0.222	0.197	0.935	0.523	0.405
Mean	0.010	0.267	0.067	0.102	0.127	0.861	0.482	0.351
St.Dev.	0.014	0.435	0.091	0.141	0.171	0.087	0.046	0.056

Naïve Bayes+filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.088	0.300	0.500	0.375	0.328	0.821	0.465	0.330
2	0.241	0.174	0.667	0.276	0.246	0.742	0.212	0.056
3	0.000	0.000	0.000	0.000	0.000	0.837	0.298	0.133
4	0.000	0.000	0.000	0.000	0.000	0.729	0.256	0.100
5	0.050	0.200	0.167	0.182	0.127	0.760	0.419	0.277
Mean	0.076	0.135	0.267	0.167	0.140	0.778	0.330	0.179
St.Dev.	0.099	0.132	0.303	0.167	0.147	0.048	0.108	0.118

Bayes Net+filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	0.000	0.000	0.000	0.000	0.932	0.570	0.465
2	0.000	0.000	0.000	0.000	0.000	0.882	0.553	0.435
3	0.000	0.000	0.000	0.000	0.000	0.966	0.500	0.366
4	0.000	0.000	0.000	0.000	0.000	0.856	0.395	0.251
5	0.050	0.200	0.167	0.182	0.127	0.878	0.419	0.277
Mean	0.010	0.040	0.033	0.036	0.025	0.903	0.487	0.359
St.Dev.	0.022	0.089	0.075	0.081	0.057	0.045	0.078	0.094

Bagging +filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.000	0.000	0.000	0.000	0.000	0.784	0.477	0.345
2	0.051	0.000	0.000	0.000	0.061	0.674	0.459	0.319
3	0.051	0.333	0.400	0.364	0.321	0.943	0.488	0.353
4	0.013	0.500	0.167	0.250	0.261	0.926	0.523	0.411
5	0.038	0.250	0..167	0.200	0.156	0.923	0.523	0.404
Mean	0.031	0.217	0.142	0.163	0.160	0.850	0.494	0.366
St.Dev.	0.023	0.217	0.189	0.160	0.134	0.117	0.028	0.040

JRIP + filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.075	0.000	0.000	0.000	0.075	0.691	0.465	0.323
2	0.000	0.000	0.000	0.000	0.000	0.781	0.424	0.219
3	0.000	0.000	0.000	0.000	0.000	0.904	0.405	0.249
4	0.000	0.000	0.000	0.000	0.000	0.677	0.465	0.286
5	0.000	1.000	0.167	0.286	0.396	0.915	0.547	0.425
Mean	0.015	0.200	0.033	0.057	0.094	0.794	0.461	0.300
St.Dev.	0.034	0.447	0.075	0.128	0.172	0.113	0.055	0.080

FURIA+ filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.100	0.200	0.333	0.250	0.185	0.678	0.453	0.321
2	0.000	0.000	0.000	0.000	0.000	0.500	0.341	0.179
3	0.000	0.000	0.000	0.000	0.000	0.500	0.476	0.333
4	0.000	0.000	0.000	0.000	0.000	0.477	0.453	0.341
5	0.013	0.750	0.500	0.600	0.590	0.792	0.488	0.340
Mean	0.023	0.190	0.167	0.170	0.155	0.590	0.442	0.297
St.Dev.	0.044	0.325	0.236	0.264	0.256	0.140	0.059	0.067

XGBoost +filter	FPR	Prec	Recall	F1-measure	MCC	AUC	Acc	Kappa
1	0.013	0.667	0.333	0.444	0.445	0.931	0.558	0.454
2	0.025	0.333	0.167	0.222	0.196	0.781	0.459	0.324
3	0.038	0.500	0.600	0.545	0.516	0.952	0.560	0.449
4	0.088	0.222	0.333	0.267	0.205	0.875	0.570	0.468
5	0.013	0.500	0.167	0.250	0.261	0.933	0.593	0.499
Mean	0.035	0.444	0.320	0.346	0.325	0.894	0.548	0.439
St.Dev.	0.031	0.171	0.177	0.142	0.147	0.070	0.052	0.067

Appendix J

Rules generated using JRIP and FURIA

JRIP rules:

===========

(NWC/TE >= 36.08) and (EarRev/OperRev >= 16.98) => Class_1=1 (7.0/0.0)

(TEPerCapita >= 4809) and (RealGDP >= 1.54) and (RealGDP <= 2.6) and (Debt/GDP <= 20.32) => Class_1=1 (8.0/0.0)

(OwnRev/OperRev <= 4.3) and (LongDebt/Debt >= 4.78) and (RealGDP >= 0) => Class_1=3 (17.0/0.0)

(OperBalance/OR >= 16.44) and (GovTransf/OperRev >= 62.9) and (RealGDP <= 1.36) => Class_1=3 (9.0/1.0)

(OperBalance/OR >= 25.48) and (MaturityDebt <= 0) => Class_1=3 (6.0/1.0)

(CashSurp/TR <= -4.44) and (LongDebt/Debt <= 0) => Class_1=6 (11.0/2.0)

(CashSurp/TR <= -2.64) and (TEPerCapita <= 967.4) and (OwnRev/OperRev <= 31.3) and (Unemployment <= 5.86) => Class_1=6 (12.0/1.0)

(TEPerCapita >= 5701) and (OwnRev/OperRev <= 86.58) => Class_1=2 (22.0/1.0)

(OwnRev/OperRev >= 71.76) and (CapitalSpend >= 19.18) => Class_1=2 (10.0/1.0)

(NationalUnemployment <= 3.42) => Class_1=2 (2.0/0.0)

(DebtSer/TR >= 26.5) => Class_1=2 (7.0/3.0)

(RealGDP >= 5.2) => Class_1=4 (9.0/1.0)

(NationalUnemployment >= 9.7) and (LongDebt/Debt >= 19.36) => Class_1=4 (24.0/6.0)

(AccualFinancingSurplus/TR >= 0.22) => Class_1=4 (5.0/1.0)

=> Class_1=5 (107.0/40.0)

Number of Rules : 15

FURIA rules:

===========

(NWC/TE in [30, 36.08, inf, inf]) and (EarRev/OperRev in [8.48, 16.98, inf, inf]) => Class_1=1 (CF = 0.79)

(TEPerCapita in [3939.75, 4809, inf, inf]) and (RealGDP in [1.42, 1.54, inf, inf]) and (Debt/GDP in [-inf, -inf, 20.32, 23.46]) and (Inter/OperRev in [1.64, 2.08, inf, inf]) => Class_1=1 (CF = 0.83)

(EarRev/OperRev in [-inf, -inf, 10.75, 10.82]) and (NationalUnemployment in [-inf, -inf, 7.18, 14.36]) and (DebtSer/TR in [13.08, 13.48, inf, inf]) and (Unemployment in [3.28, 4.3, inf, inf]) => Class_1=2 (CF = 0.92)

(TEPerCapita in [7435.8, 7888.8, inf, inf]) and (ShortDebt/Debt in [9.86, 10.32, inf, inf]) and (RealGDP in [-inf, -inf, 1.42, 2.28]) => Class_1=2 (CF = 0.85)

(TEPerCapita in [1556, 1579.25, inf, inf]) and (NWC/TE in [16.32, 17.94, inf, inf]) => Class_1=2 (CF = 0.71)

(EarRev/OperRev in [-inf, -inf, 4.58, 6.3]) and (Inter/OperRev in [-inf, -inf, 1.92, 1.96]) and (Unemployment in [-inf, -inf, 5.08, 5.15]) and (GovTransf/OperRev in [10.45, 13.42, inf, inf]) => Class_1=2 (CF = 0.81)

(OwnRev/OperRev in [-inf, -inf, 4.3, 4.66]) and (LongDebt/Debt in [0, 4.78, inf, inf]) and (RealGDP in [-0.38, 0, inf, inf]) => Class_1=3 (CF = 0.91)

(OperBalance/OR in [12.9, 16.44, inf, inf]) and (GovTransf/OperRev in [44.48, 62.9, inf, inf]) and (RealGDP in [-inf, -inf, 1.36, 2.4]) => Class_1=3 (CF = 0.76)

(OperBalance/OR in [17.53, 25.48, inf, inf]) and (Debt/GDP in [1.78, 2.34, inf, inf]) => Class_1=3 (CF = 0.76)

(NationalUnemployment in [9.18, 9.7, inf, inf]) and (LongDebt/Debt in [13.38, 19.36, inf, inf]) and (Unemployment in [-inf, -inf, 20.04, 20.86]) => Class_1=4 (CF = 0.83)

(RealGDP in [5.18, 5.2, inf, inf]) and (NationalUnemployment in [-inf, -inf, 5.18, 5.52]) => Class_1=4 (CF = 0.83)

(NationalUnemployment in [4.84, 11, inf, inf]) and (EarRev/OperRev in [72.6, 79.5, inf, inf]) => Class_1=4 (CF = 0.65)

(NationalUnemployment in [27.22, 29.8, inf, inf]) and (Debt/GDP in [0.18, 0.32, inf, inf]) => Class_1=4 (CF = 0.62)

(DebtSer/TR in [-inf, -inf, 0.56, 0.58]) and (GovTransf/OperRev in [79.98, 91.08, inf, inf]) => Class_1=4 (CF = 0.7)

(TEPerCapita in [-inf, -inf, 1418.6, 1456]) and (TE/GDP in [5.8, 5.9, inf, inf]) and (TEPerCapita in [996.6, 1056, inf, inf]) => Class_1=5 (CF = 0.88)

(TEPerCapita in [-inf, -inf, 947.2, 967.4]) and (RealGDP in [1.58, 1.63, inf, inf]) and (RealGDP in [-inf, -inf, 3.63, 4]) => Class_1=5 (CF = 0.63)

(AccualFinancingSurplus/TR in [-inf, -inf, -11.65, -10.4]) => Class_1=5 (CF = 0.68)

(CashSurp/TR in [-inf, -inf, -4.12, -3.73]) and (LongDebt/Debt in [-inf, -inf, 0, 0.1]) => Class_1=6 (CF = 0.73)

(TEPerCapita in [-inf, -inf, 612.4, 785.8]) and (OwnRev/OperRev in [-inf, -inf, 28.42, 30.64]) and (Unemployment in [-inf, -inf, 5.86, 5.88]) => Class_1=6 (CF = 0.79)

Number of Rules : 19