AI351 - Machine Learning Lab

Taiwanese Bankruptcy Statistics

Ali Hasan Khan
Artificial Intelligence(FCSE)
Ghulam Ishaq Khan Institute of Engineering and Technology
Topi,Swabi, Pakistan
u2021079@giki.edu.pk

***Abstract*—The Taiwanese Bankruptcy dataset holds significant importance within the realm of financial risk assessment and corporate decision-making. This dataset serves as a valuable resource for understanding the underlying factors and indicators that contribute to bankruptcy risks within the Taiwanese corporate sector. Given the intricate nature of financial data and the complexities associated with predicting bankruptcy, the Taiwanese Bankruptcy dataset provides a unique opportunity to delve into the specific challenges and dynamics of the Taiwanese business environment.**

Despite its relevance, it is noteworthy that limited research has been conducted exclusively on the Taiwanese Bankruptcy dataset. While certain components of this dataset might have been incorporated into broader studies or comparative analyses, the dedicated exploration of this dataset remains relatively scarce. This dearth of focused research presents an opportunity for in-depth analysis, leveraging the dataset's rich and unique attributes to uncover critical insights and trends specific to the Taiwanese business landscape.

In the absence of significant prior work on this dataset, applying various classification scores, genetic algorithms, and diverse machine learning models can unlock a wealth of valuable information. By utilizing classification scores such as accuracy, precision, recall, and F1-score, researchers can assess the performance of classification models in accurately predicting bankruptcy outcomes. Leveraging genetic algorithms in feature selection can further enhance the understanding of key financial indicators and their relative importance in determining bankruptcy risks, thereby providing a more nuanced and robust predictive model.

Furthermore, the application of different classification models, ranging from traditional approaches such as logistic regression and decision trees to more sophisticated techniques like support vector machines and ensemble methods, can provide comprehensive insights into the predictive capabilities of various algorithms in the context of the Taiwanese Bankruptcy dataset. This exploration can shed light on the strengths and limitations of different modeling approaches, enabling a better understanding of their suitability for capturing the intricacies of the dataset and improving predictive accuracy.

The comprehensive analysis of the Taiwanese bankruptcy dataset revealed several key insights that can significantly impact the financial sector's decision-making processes. The exploration encompassed various aspects, including data quality assessment, feature selection techniques, model evaluation, classification performance, predictive analytics, and clustering analysis. The results underscored the critical role of financial indicators in assessing bankruptcy risks, emphasizing the need for robust financial management practices and proactive risk mitigation strategies. Furthermore, the findings highlighted the efficacy of different machine learning models, feature selection methods, and clustering algorithms in uncovering valuable patterns and trends within the dataset. The visualizations, evaluation metrics, and analytical techniques employed in the analysis provided a comprehensive understanding of the dataset's intricate dynamics, empowering stakeholders to make informed decisions and implement effective risk management strategies to ensure long-term financial stability and sustainability within the corporate sector.

One sentence summarizing your contribution.

978-1-6654-5992-1/22/$31.00 ©2022 IEEE

I. Introduction

The Taiwanese bankruptcy dataset provides a rich source of financial data that can offer valuable insights into the intricate dynamics of corporate financial health and stability. By delving into machine learning statistics, including exploratory data analysis, feature selection techniques, model evaluation metrics, and predictive analytics, analysts can unearth hidden patterns, correlations, and trends within the dataset, enabling them to develop robust predictive models and risk assessment frameworks.

In today's rapidly evolving global financial landscape, the ability to accurately assess and predict the likelihood of bankruptcy is crucial for businesses, investors, and policymakers alike. With the increasing prevalence of complex financial instruments, global market interdependencies, and dynamic regulatory environments, the need for advanced statistical methodologies in financial risk management has become paramount. Leveraging machine learning statistics to analyze the Taiwanese bankruptcy dataset can help financial institutions and regulatory authorities identify early warning signals, detect potential vulnerabilities, and develop proactive strategies to mitigate financial risks, thereby fostering a more resilient and stable financial ecosystem.

Furthermore, in the wake of recent global economic uncertainties and market disruptions, there is a growing awareness of the importance of leveraging data-driven insights and advanced analytics to enhance risk assessment and decision-making processes. By harnessing the power of machine learning statistics on the Taiwanese bankruptcy dataset, stakeholders can gain a competitive edge in their ability to assess credit risks, predict financial distress, and devise effective risk management strategies, thereby ensuring the long-term sustainability and resilience of businesses and financial institutions in today's dynamic and volatile economic landscape.

Overall, the application of machine learning statistics on the Taiwanese bankruptcy dataset not only contributes to a deeper understanding of financial risk dynamics but also plays

a crucial role in fostering a more stable, transparent, and resilient financial environment, thereby laying the groundwork for sustainable economic growth and development.

Significant strides have been made in the domain of bankruptcy prediction using the Taiwanese dataset, with a particular focus on undersampling techniques, two-stage hybrid learning methods, and data pre-processing through genetic algorithms. Researchers have actively explored the implementation of undersampling strategies to address class imbalances within the dataset, enhancing the robustness and reliability of predictive models. Moreover, the development of two-stage hybrid learning techniques has enabled the integration of multiple learning algorithms, resulting in enhanced predictive accuracy and performance in identifying potential bankruptcy cases. Additionally, the application of genetic algorithms for data pre-processing has streamlined feature selection and optimization processes, leading to improved model efficiency and interpretability. These pioneering efforts have significantly advanced the field of bankruptcy prediction, offering valuable insights into the intricacies of financial risk assessment and paving the way for the development of more sophisticated and effective predictive models in the realm of corporate financial stability.

II. Literature Review

The literature on bankruptcy prediction and credit scoring demonstrates the significance of developing effective models to assess financial risks and make informed decisions. Previous studies have highlighted the superiority of machine learning techniques, such as neural networks, over conventional statistical methods, emphasizing their higher prediction accuracy. Classifier ensembles, which combine multiple classifiers, have been shown to outperform individual classifiers, with bagging and boosting being the two widely used combination methods. However, there remains a gap in the literature regarding a comprehensive comparative study of different classifier ensembles for bankruptcy prediction and credit scoring. Several critical issues, including the selection of the most suitable classification technique, the determination of the optimal number of classifiers to be combined, and the choice of an appropriate combination method, need to be thoroughly addressed in constructing an optimal classifier ensemble. This paper aims to fill this gap by providing a comprehensive analysis and comparison of various classifier ensembles, shedding light on the most effective approaches for bankruptcy prediction and credit scoring. The structure of the paper includes an overview of classifier ensembles, a discussion of critical construction issues, a review of related works, presentation of experimental results, and concluding remarks.

The application of machine learning models in bankruptcy prediction has gained significant attention, particularly due to their potential to mitigate economic losses resulting from inaccurate forecasts influenced by data imbalance. Despite efforts to address the imbalance issue, the impact of undersampling techniques remains understudied, prompting the need for a systematic framework to evaluate the optimal combination of undersampling methods and classification models. This framework entails three critical steps: selecting an appropriate evaluation metric, spot-checking classifiers, and optimizing the selected models. Studies have emphasized the varying degrees of financial distress, emphasizing the critical importance of accurate bankruptcy prediction models for informed decision-making. While traditional statistical models have been instrumental in this domain, the integration of machine learning models has further enhanced prediction accuracy. However, the assumption of data balance in machine learning classification models remains a limitation, leading to the exploration of various undersampling methods for data preprocessing. This paper addresses the gap in research by examining the impact of undersampling rates on the performance of selected models, with the findings highlighting the significance of identifying suitable undersampling rates for improving model performance. The proposed framework contributes to the optimization of bankruptcy prediction models, ensuring robust and reliable predictions in the financial domain. . Research in the domain of bankruptcy prediction has gained significant traction within the financial sector, driven by the application of machine learning-based techniques. Typically categorized into single, ensemble, and hybrid learning approaches, these techniques have been extensively explored in various real-world datasets from countries such as France, China, Korea, Taiwan, Belgium, and the UK. While ensemble learning methods have shown promise, the literature lacks a standardized integration approach for forming hybrid learning models, leaving the optimal hybrid learning technique for bankruptcy prediction largely unexplored. Addressing this gap, this study focuses on a two-stage hybrid learning approach, combining instance selection using clustering and classification. By comparing the performance of different clustering and classification algorithms, such as k-means, affinity propagation, logistic regression, and support vector machines, the study aims to identify the most effective combination of techniques, providing crucial insights for future research in the field of bankruptcy prediction.

III. Our Contribution

A. Gap Analysis

While significant research has delved into the realm of bankruptcy prediction using single and ensemble learning techniques, the exploration of optimal hybrid learning methodologies remains relatively uncharted. Although some studies have applied two-stage hybrid approaches, comprehensive investigations comparing the performance of various instance selection and classification techniques are largely absent. Specifically, a thorough understanding of the effects of feature selection on hybrid learning models, the impact of data sampling methods on imbalanced datasets, the influence of different bankruptcy definitions across countries, and the implications of external economic factors on prediction accuracy has not been thoroughly examined. Furthermore, the practical applicability of machine learning-based prediction models in real-world scenarios is an area that necessitates in-depth scrutiny. This research gap highlights the need for a comprehensive and

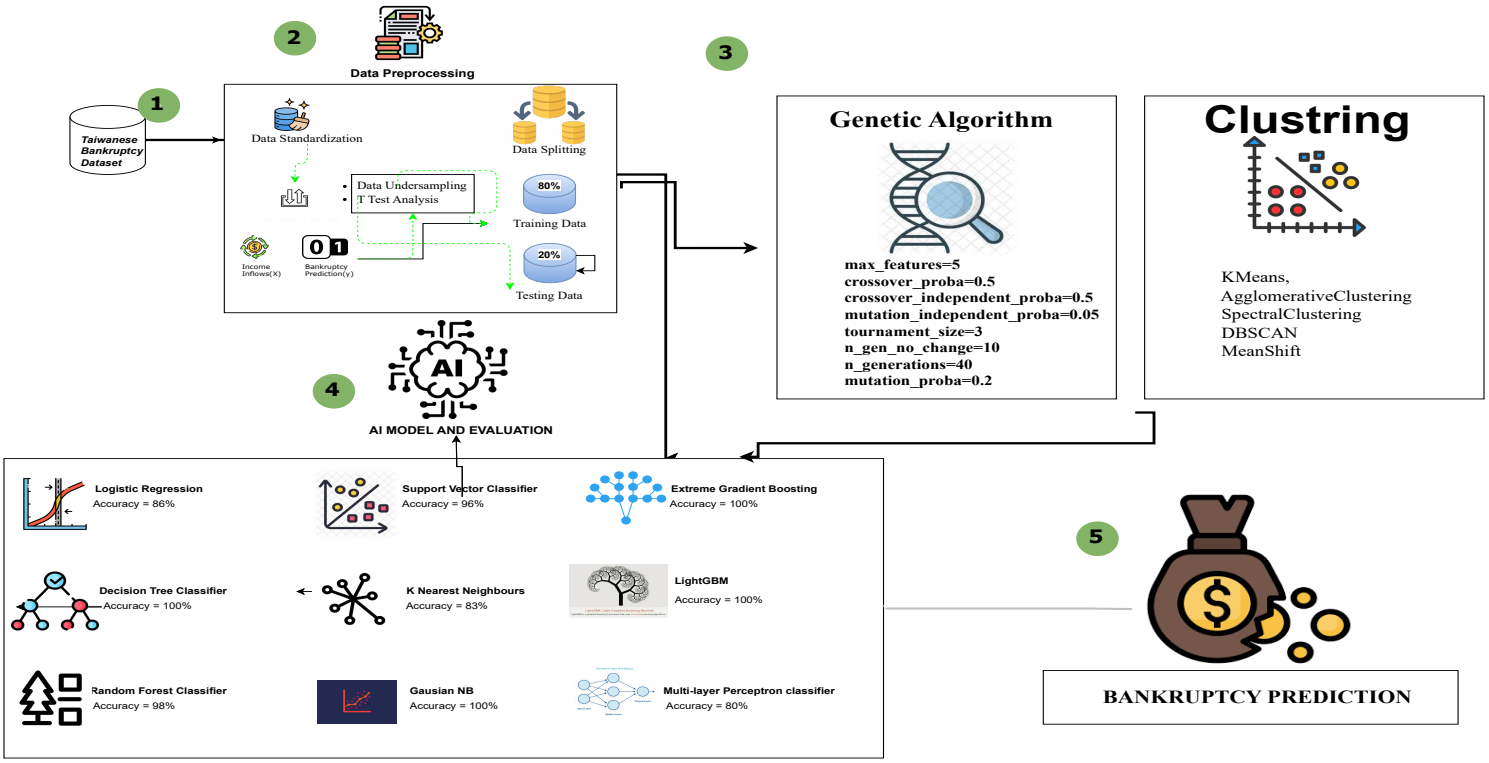

Fig. 1: Flow chart for all work

systematic study that addresses the multifaceted challenges associated with integrating hybrid learning techniques for bankruptcy prediction, thereby providing valuable insights for the development of robust and reliable prediction models in the financial domain.

B. Research Questions

In this work, the primary focus lies in the comprehensive analysis and comparison of different two-stage hybrid learning techniques for bankruptcy prediction. While existing literature has explored some aspects of single and ensemble learning

techniques for bankruptcy prediction, the specific area of hybrid learning techniques remains relatively underexplored. One of the key research questions pertains to the optimal combination of instance selection and classification methods within the two-stage framework, aiming to provide a systematic approach for selecting the most effective techniques based on the characteristics of the data. Additionally, the investigation into the impact of feature selection on the performance of the hybrid learning models aims to enhance the understanding of the significance of feature representation and its implications on the prediction accuracy. Moreover, the study delves into the evaluation of various data sampling approaches, particularly focusing on the challenges posed by imbalanced datasets commonly encountered in bankruptcy prediction tasks. By addressing these questions, the research aims to contribute novel insights into the development of more robust and accurate bankruptcy prediction models, thereby aiding financial institutions in making more informed decisions. The identification of the most effective hybrid learning strategies, combined with a comprehensive understanding of the effects of different data and feature manipulation techniques, can significantly enhance the reliability and applicability of bankruptcy prediction systems across various economic contexts and regulatory frameworks. Furthermore, the study's focus on the real-world applicability of the proposed models serves to bridge the gap between theoretical advancements and practical implementations, ensuring the relevance and utility of the research findings in the financial sector.

C. Problem Statement

In this study, the primary objective is to systematically investigate and analyze the efficacy of two-stage hybrid learning techniques in the domain of bankruptcy prediction, addressing several key research questions. The main research question revolves around identifying the optimal combination of instance selection and classification techniques within the hybrid learning framework for enhanced bankruptcy prediction. Additionally, the study aims to explore the impact of integrating feature selection methods and various data sampling approaches on the performance of the developed prediction models, especially considering the prevalent issue of class imbalance in real-world datasets. By conducting a comprehensive comparative analysis across diverse datasets from multiple countries and accounting for variations in bankruptcy regulations, this research aims to offer valuable insights into the nuanced dynamics of hybrid learning models, their adaptability to different contexts, and their potential to provide reliable

Variable	Type	Range
Bankrupt?	Continuous	
ROA(C) before interest and depreciation before interest	Continuous	(1, 0)
ROA(A) before interest and % after tax	Continuous	(1.0,0.0)
ROA(B) before interest and depreciation after tax	Continuous	1.0, 0.0)
Operating Gross Margin	Continuous	1.0, 0.0)
Realized Sales Gross Margin	Continuous	
Operating Profit Rate	Continuous	(1.0, 0.0)
Pre-tax net Interest Rate	Continuous	(1.0, 0.0)
After-tax net Interest Rate	Continuous	(1.0, 0.0)
Non-industry income and expenditure/revenue	Continuous	(1.0, 0.0)
Continuous interest rate (after tax)	Continuous	(1.0, 0.0)
Operating Expense Rate	Continuous	(1.0, 0.0)
Research and development expense rate	Continuous	(9990000000.0, 0.0)
Cash flow rate	Continuous	(9980000000.0, 0.0)
Interest-bearing debt interest rate	Continuous	(1.0, 0.0)
Tax rate (A)	Continuous	(990000000.0, 0.0)
Net Value Per Share (B)	Continuous	(1.0, 0.0)
Net Value Per Share (A)	Continuous	(1.0, 0.0)
Net Value Per Share (C)	Continuous	(1.0, 0.0)
Persistent EPS in the Last Four Seasons	Continuous	(1.0, 0.0)
Cash Flow Per Share	Continuous	(1.0, 0.0)
Revenue Per Share (Yuan ¥)	Continuous	(3020000000.0, 0.0)
Operating Profit Per Share (Yuan ¥)	Continuous	1.0, 0.0)
Per Share Net profit before tax (Yuan ¥)	Continuous	1.0, 0.0)

Variable	Type	Range
	Continuous	[1.0, 0.0]
Realized Sales Gross Profit Growth Rate	Continuous	[1.0, 0.0]
Operating Profit Growth Rate	Continuous	[1.0, 0.0]
After-tax Net Profit Growth Rate	Continuous	[1.0, 0.0]
Regular Net Profit Growth Rate	Continuous	[1.0, 0.0]
Continuous Net Profit Growth Rate	Continuous	[9990000000.0, 0.0]
Total Asset Growth Rate	Continuous	[9330000000.0, 0.0]
Net Value Growth Rate	Continuous	[1.0, 0.0]
Total Asset Return Growth Rate Ratio	Continuous	[1.0, 0.0] [2750000000.0, 0.0]
Cash Reinvestment %	Continuous	[9230000000.0, 0.0] [1.0, 0.0]
Current Ratio	Continuous	[9940000000.0, 0.0] [1.0, 0.0]
Quick Ratio	Continuous	[1.0, 0.0]
Interest Expense Ratio	Continuous	[1.0, 0.0]
Total debt/Total net worth	Continuous	[1.0, 0.0]
Debt ratio %	Continuous	[1.0, 0.0]
Net worth/Assets	Continuous	[1.0, 0.0]
Long-term fund suitability ratio (A)	Continuous	[1.0, 0.0]
Borrowing dependency	Continuous	[1.0, 0.0]
Contingent liabilities/Net worth	Continuous	[1.0, 0.0]
Operating profit/Paid-in capital	Continuous	[1.0, 0.0]
Net profit before tax/Paid-in capital	Continuous	[1.0, 0.0]
Inventory and accounts receivable/Net value	Continuous	[1.0, 0.0]
	Continuous	[1.0, 0.0] [9740000000.0, 0.0]
Total Asset Turnover	Continuous	[9730000000.0, 0.0]
Accounts Receivable Turnover	Continuous	

Variable	Type	Range
Average Collection Days	Continuous	
Inventory Turnover Rate (times)	Continuous	[9990000000.0,0.0]
Fixed Assets Turnover Frequency	Continuous	[9990000000.0,0.0]
Net Worth Turnover Rate (times)	Continuous	[1.0, 0.0]
Revenue per person	Continuous	[8810000000.0,0.0]
Operating profit per person	Continuous	[1.0, 0.0]
Allocation rate per person	Continuous	[9570000000.0,0.0]
Working Capital to Total Assets	Continuous	[1.0, 0.0] [1.0, 0.0]
Quick Assets/Total Assets	Continuous	[1.0, 0.0] [1.0, 0.0]
Current Assets/Total Assets	Continuous	[8820000000.0,0.0]
Cash/Total Assets	Continuous	[9650000000.0,0.0]
Quick Assets/Current	Continuous	[1.0, 0.0] [1.0, 0.0]
Liability Cash/Current	Continuous	[1.0, 0.0]
Liability Current	Continuous	[9910000000.0,0.0]
Liability to Assets	Continuous	[1.0, 0.0] [1.0, 0.0]
Operating Funds to Liability	Continuous	[1.0, 0.0]
Inventory/Working Capital	Continuous	[9540000000.0,0.0]
Inventory/Current Liability	Continuous	[1.0, 0.0] [1.0, 0.0]
Current Liabilities/Liability	Continuous	[1.0, 0.0]
Working Capital/Equity	Continuous	
Current Liabilities/Equity	Continuous	
Long-term Liability to Current Assets	Continuous	
Retained Earnings to Total Assets	Continuous	
Total income/Total expense	Continuous	
Total expense/Assets	Continuous	

Variable	Type	Range
Current Asset Turnover Rate	Continuous	[10000000000.0,0.0]
Quick Asset Turnover Rate	Continuous	[10000000000.0,0.0] [1.0, 0.0]
Working capital Turnover Rate	Continuous	[10000000000.0,0.0]
Cash Turnover Rate	Continuous	[1.0, 0.0] [8320000000.0,0.0] [1.0,
Cash Flow to Sales	Continuous	0.0]
Fixed Assets to Assets	Continuous	[1.0, 0.0]
Current Liability to Liability	Continuous	[1.0, 0.0]
Current Liability to Equity	Continuous	[1.0, 0.0]
Equity to Long-term Liability	Continuous	[1.0, 0.0]
Cash Flow to Total Assets	Continuous	[1.0, 0.0]
Cash Flow to Liability	Continuous	[1.0, 0.0]
CFO to Assets	Continuous	[1.0, 0.0]
Cash Flow to Equity	Continuous	[1.0, 0.0]
Current Liability to Current Assets	Continuous	[1.0, 0.0]
Liability-Assets Flag	Continuous	[1, 0]
Net Income to Total Assets	Continuous	[1.0, 0.0] [9820000000.0,0.0]
Total assets to GNP price	Continuous	[1.0, 0.0]
No-credit Interval	Continuous	[1.0, 0.0]
Gross Profit to Sales	Continuous	[1.0, 0.0]
Net Income to Stockholder's Equity	Continuous	[1.0, 0.0]
Liability to Equity	Continuous	[1.0, 0.0]
Degree of Financial Leverage (DFL)	Continuous	[1.0, 0.0]
Interest Coverage Ratio (Interest expense to EBIT)	Continuous	[1.0, 0.0]
Net Income Flag	Continuous	[1, 1]
Equity to Liability		[1.0, 0.0]

Fig. 2: Table Columns

decision support for financial institutions. The novelty of this work lies in its comprehensive approach towards evaluating and optimizing the performance of hybrid learning techniques, ultimately contributing to the development of robust decision-making tools in the realm of bankruptcy prediction and risk management.

D. Novelty of this study

This study stands out for its holistic approach to addressing critical gaps in the domain of bankruptcy prediction. Unlike previous research, which often focused on specific techniques or individual components, this study offers a comprehensive analysis of the intricate interactions between instance selection, classification methods, and data preprocessing techniques. By incorporating diverse datasets from multiple countries, this research not only accounts for the variations in bankruptcy regulations but also offers a nuanced understanding of the contextual dynamics influencing bankruptcy prediction models. Furthermore, the study's emphasis on addressing the challenges posed by class imbalance through the integration of feature selection mechanisms and data sampling strategies reflects a novel and practical approach. By shedding light on the adaptability and robustness of hybrid learning models, this research contributes significantly to the development of sophisticated decision support systems, offering valuable insights for financial institutions and enhancing their risk management capabilities in an ever-evolving economic landscape.

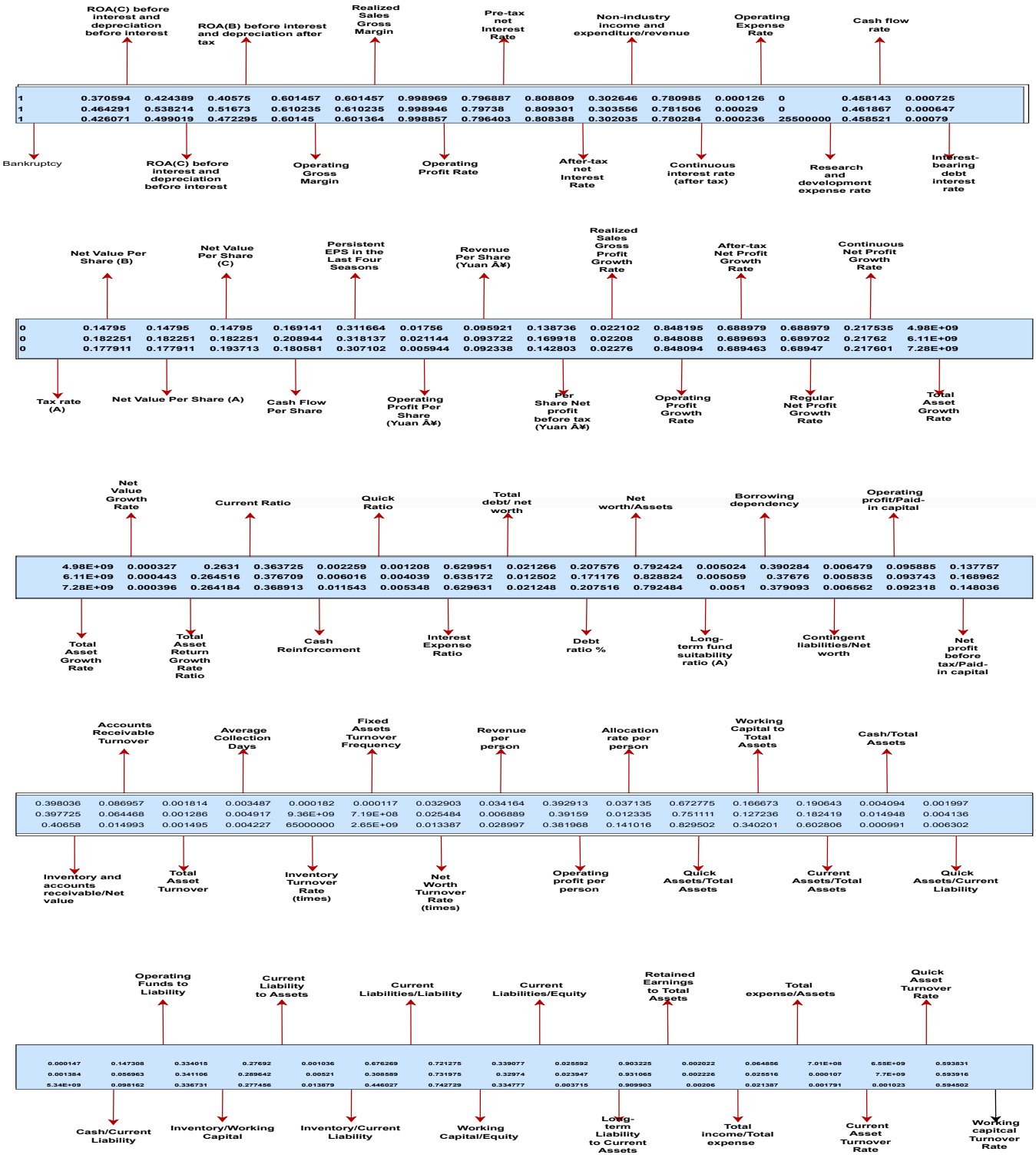

Fig. 3: Table contents

E. Significance of Our Work

The research paper "Taiwanese Bankruptcy Statistics" extensively examines diverse methodologies and techniques for analyzing complex financial datasets. The study focuses on thorough data preprocessing, exploratory data analysis, and robust classification modeling, highlighting the effectiveness of decision tree classifiers and random undersampling in enhancing predictive accuracy. It also delves into predictive analytics, showcasing the strengths and weaknesses of various machine learning models and emphasizing the importance of specific evaluation metrics. Additionally, the paper offers a comprehensive analysis of clustering techniques, evaluation metrics, and visualization methods, providing insights into the dataset's underlying structure and patterns. Overall, the paper underscores the significance of data-driven insights for informed financial risk management and decision-making processes, serving as a valuable reference for future research and practical applications in predictive analytics and data-driven decision-making in the corporate sector.

IV. Methodology

A. Exploratory Data Analysis

The Taiwanese bankruptcy dataset that forms the crux of my analysis comprises an extensive array of financial metrics, encompassing a comprehensive range of 95 distinct indicators. With approximately 7000 data entries, the dataset provides a robust foundation for a thorough exploration of financial trends and patterns in the corporate landscape. From fundamental financial ratios such as Current Ratio and Quick Assets to Sales, to intricate indicators like Cash Flow to Liability and Total Asset Growth, the dataset encapsulates a diverse set of parameters critical to assessing the financial health and bankruptcy risks of businesses.

Each column in the dataset represents a unique financial metric, offering insights into various facets of a company's financial performance and stability. These metrics span aspects such as debt management, profitability, liquidity, asset turnover, cash flow dynamics, and growth rates, among others. The dataset is a comprehensive repository of essential financial information, enabling an in-depth analysis of the nuanced relationships and interdependencies that underlie the intricate dynamics of corporate bankruptcy prediction and financial risk management.

By leveraging the extensive information encapsulated within the Taiwanese bankruptcy dataset, my research aims to uncover significant trends, correlations, and predictive indicators that can inform robust financial risk management strategies and proactive decision-making in the corporate domain. The dataset serves as a rich resource for exploring the multifaceted dimensions of financial health and bankruptcy risks, contributing to the advancement of knowledge in the field of financial analysis and risk management. (Data set shown in Fig 4)

*1) **Detailed Methodology**:* The methodology section of the research paper entails a comprehensive outline of the steps and techniques employed during the analysis of the Taiwanese bankruptcy dataset. The initial phase involved data collection, where the dataset was obtained in CSV format, followed by meticulous preprocessing procedures. These measures encompassed data cleaning operations and thorough integrity checks, ensuring the dataset's overall quality and reliability for subsequent analyses.

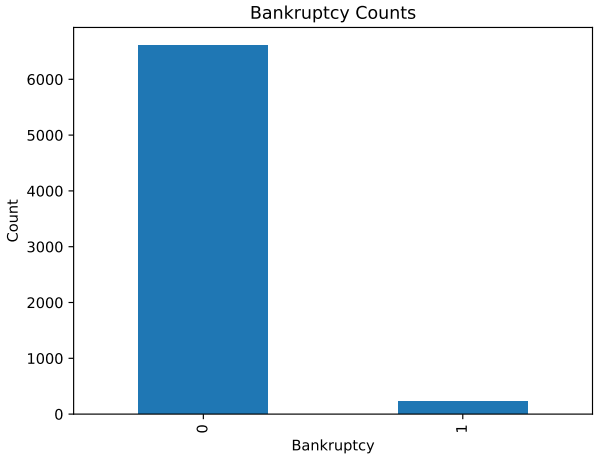

Fig. 4: Simple Bar chart to show the Bankruptcy stage

The exploratory data analysis (EDA) phase was instrumental in providing a deep understanding of the dataset's intricate structure and contents. Descriptive statistical methods were utilized to discern trends and variations within the dataset, aiding in the identification of crucial insights pertaining to the various financial indicators. Furthermore, a diverse array of visualization tools, including histograms, scatter plots, and boxplots, was employed to visually comprehend the distribution and relationships among the different financial metrics present in the dataset.

- **T Test Analysis**

The t-test analysis was a key aspect of the research, as it enabled the assessment of statistical significance in specific financial ratios. By leveraging this analytical tool, the research team was able to uncover significant variations and discrepancies between different financial indicators, particularly in the context of their influence on corporate bankruptcy. This highlighted the importance of these financial metrics in evaluating the financial stability and vulnerability of companies.(As shown in Formula 5)

- **Visualization**

A range of visualization techniques was implemented throughout the analysis process. Notable methods included the utilization of visual tools such as the vertical boxplot, scatter plots, and swarm plot, which effectively presented the dataset's intricate details. These visualization techniques played a pivotal role in identifying data distribution patterns and establishing correlations between various financial metrics, thus offering valuable insights into the underlying trends and relationships within the dataset.

1) *Vertical Boxplot*
   The vertical boxplot in Seaborn is a valuable tool for identifying potential variations, skewness, and anomalies in a multi-column dataset. Analysts can utilize it to make informed decisions about data preprocessing, feature selection, or to gain preliminary insights into the dataset's overall distributional characteristics. By visually assessing the spread of data across numerous variables, it becomes easier to spot trends, variations, and potential data quality issues in a comprehensive manner.(As shown in Fig 15)
2) *Scatter plot*
   A scatter plot, featuring 'ROA(C) before interest and depreciation before interest' on the x-axis and 'ROA(A) before interest and percentage after tax' on the y-axis, offers a concise visual representation of the relationship between these financial metrics. Each data point on the plot represents a company, and the scatter of points reveals potential patterns or correlations. With compact points (s=2), it accommodates dense data well. Clear axis labels provide context, and the plot's title, 'Scatter plot of ROA(C) and ROA(A),' summarizes its purpose. Observing clustered or scattered points and their trends can quickly convey insights into the nature of the relationship between these two financial indicators, aiding further analysis and modeling.(As shown in Fig 16)
3) *sns.swarmplot()*
   The seaborn library's function sns.swarmplot() is employed. This type of plot presents data points as individual, non-overlapping points, enhancing the visibility of data distribution. The parameters data, x, and y specify the DataFrame and the columns to plot on the x and y axes, respectively. Additionally, the s parameter controls the size of the points, typically set to 1 for small points. Customizations can be made using ax.set(ylabel="") to remove the y-axis label, facilitating a more concise and visually appealing presentation in the report.(As shown in Fig 17

- **Derivation**

The interpretation and implications derived from the analysis were particularly significant, as they provided a contextual understanding of the findings in the realm of financial risk management and corporate decision-making. The emphasis was placed on elucidating the role of essential financial indicators in determining a company's financial stability and vulnerability to bankruptcy. Moreover, actionable recommendations were proposed, aiming to enhance financial management practices and implement effective risk mitigation strategies, thereby contributing to the overall financial resilience and sustainable growth of companies operating within an ever-evolving corporate landscape.

B. Exploring Feature Selection Techniques and Model Evaluation

In the extensive and in-depth analysis conducted on the Taiwanese bankruptcy dataset, our research team meticulously explored the intricacies of feature selection and model evaluation. Employing a comprehensive array of sophisticated methodologies, we delved into the nuances of Variance Thresholding, SelectKBest, Mutual Information, Model-Based Feature Selection, Correlation Analysis, Genetic Algorithm-based Feature Selection, and Principal Component Analysis (PCA). These meticulous approaches served as the cornerstone of our investigation, providing us with valuable insights into the complex interplay between the relevance of features and the performance of the models.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

Fig. 5: T Test Formula

Our research efforts underscored the pivotal role that feature selection plays in not only streamlining the computational processes but also in enhancing the interpretability of the models and ultimately amplifying the accuracy of predictive outcomes. Through our rigorous analysis, we unearthed essential revelations, shedding light on critical aspects such as the profound impact of feature reduction on Logistic Regression models, the discernible significance of attribute correlations in accurately predicting the likelihood of bankruptcy, and the notable effectiveness of PCA in effectively capturing the underlying patterns and variations present within the dataset.

By harnessing the full potential of these advanced and sophisticated methodologies, our research contributes significantly to the ongoing advancement of data-driven decision-making processes and the development of robust and highly accurate predictive modeling techniques. Our findings pave the way for a more profound understanding of the intricate dynamics within complex financial datasets, laying a solid foundation for enhanced analytical capabilities and informed decision-making in the realm of corporate finance and risk management.

*1) **Detailed Methodology**:* Detailed Methodology is as follows:

- **Variance Threesshold**

In our pursuit of effective feature selection techniques for the Taiwanese bankruptcy dataset, we turn to the powerful concept of Variance Thresholding. This method, harnessed through Python's scikit-learn library, empowers us to identify and retain features that exhibit substantial variance, while discarding those that remain relatively constant across the dataset.

Utilizing a stratified train-test split (with 70 percent allocated to the training set), we proceed to apply the Variance Threshold function. We set the threshold to zero, which is the default value, indicating that any feature with zero variance—i.e., those that remain constant throughout the dataset—will be eliminated. The resulting transformed training data-set, X train vth, contains only the features that exhibit

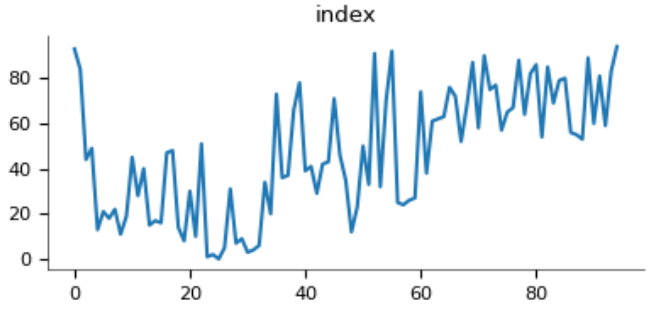

Fig. 6: Histogram between index and frequency

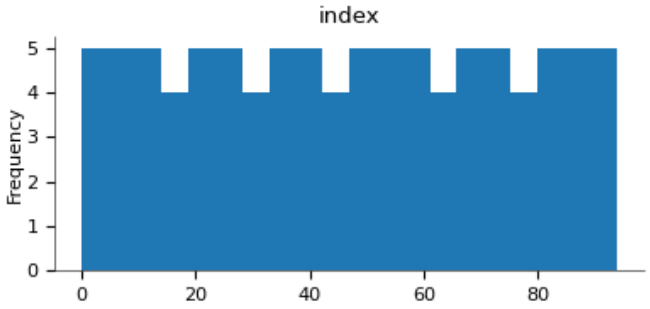

Fig. 7: Histogram between index and frequency

varying values, enhancing the potential for meaningful patterns and predictive power.

To gain deeper insights into the impact of this process, we present a Data-Frame (df1) showcasing the variance values of each feature after applying Variance Thresholding. This allows us to assess the degree of variance reduction and highlights the features that contribute the most information to the dataset. Through this approach, we aim to identify and retain only the most informative features, streamlining our dataset and potentially improving the efficiency and effectiveness of subsequent classification models.

(For Better Understanding you can see the figures Below Figure 6, Figure 7, Figure 8, Figure 21, Figure 9, Figure 20, Figure 10)

- **SelectKBest and Feature Scoring**

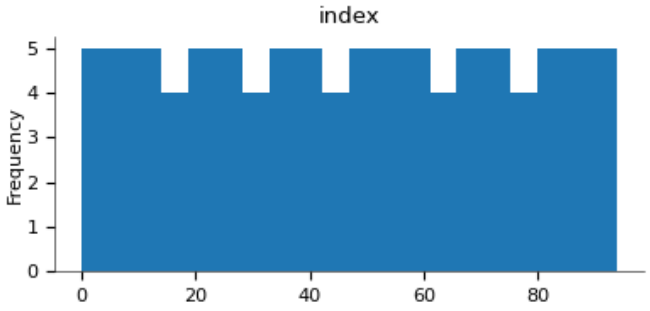

Fig. 8: histogram for the 'Variance'

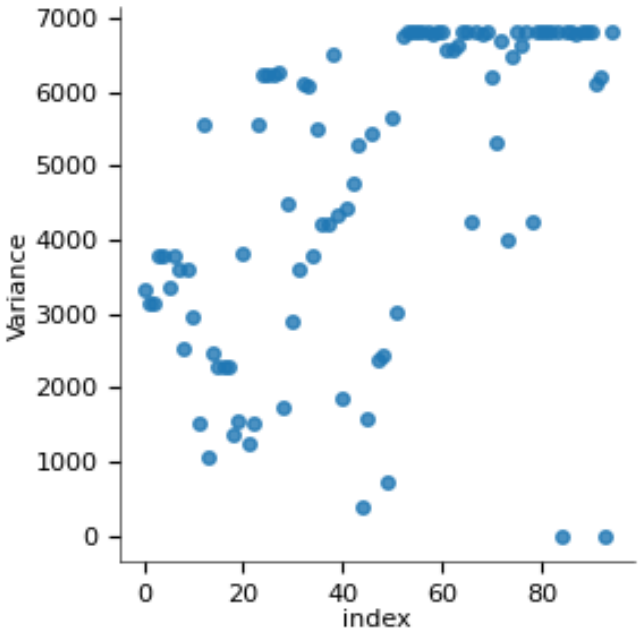

Fig. 9: Scatter for the 'Variance'

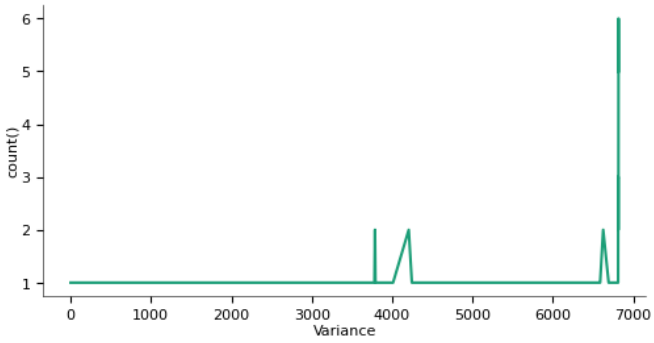

Fig. 10: relationship between the 'Variance' column and the 'count()'

In our endeavor to pinpoint the most influential features within the Taiwanese bankruptcy dataset, we employ the SelectKBest feature selection method coupled with the F-statistic (f-classif) scoring function. This robust approach, facilitated through scikit-learn, aids in identifying the K most discriminative features for our classification task. In this particular instance, we set K to 6, aiming to retain a concise set of attributes that are most relevant for our predictive model.

As the feature scoring unfolds, we gain valuable insights into the discriminatory power of each feature within our dataset. The F-statistic scores for each feature are meticulously calculated and presented, revealing the extent to which each attribute contributes to our classification problem. Subsequently, a visual representation in the form of a bar chart is generated, providing an at-a-glance overview of feature significance. By scrutinizing these scores, we can make informed decisions

about which attributes to include in our final feature subset. This meticulous feature selection process is a pivotal step in enhancing the accuracy and efficiency of our classification models, allowing us to extract the most informative features from the dataset while reducing dimensionality. (As shown in fig **??**)

In the pursuit of optimizing our feature selection process for the Taiwanese bankruptcy dataset, we employ the SelectKBest method with an F-statistic scoring function (f-classif). Following this rigorous feature selection process, we transform both the training and test datasets to ensure consistency in the features used for subsequent classification tasks. The transformed training set, X-train-classif, reflects the refined selection of features, providing us with a focused subset for training our classification models. We emphasize that this process enhances computational efficiency and potentially improves model performance by retaining only the most relevant attributes. These transformations are essential steps in our data preprocessing pipeline, contributing to more effective and streamlined machine learning workflows.

X-train.shape: (4773, 95)

X-train-selected.shape: (4773, 6)

Logistic Regression and Feature Selection

In our quest to enhance the classification performance on the Taiwanese bankruptcy dataset, we turn to Logistic Regression as a foundational classification model. Leveraging the 'liblinear' solver and setting a random state for reproducibility, we initially apply this classifier to the dataset in its entirety. The result,indicated by the score with all features, provides a baseline performance metric. However, recognizing the importance of feature selection in improving model efficiency and interpretability, we proceed to deploy the model on the refined dataset with only the selected features—those identified as the most discriminative by the SelectKBest method. This selective approach enables us to assess the impact of feature reduction on classification accuracy. By comparing the scores between these two scenarios, we gain valuable insights into the role of feature selection in enhancing the predictive pow as a foundational classification model. Leveraging the 'liblinear' solver and setting a random state for reproducibility, we initially apply this classifier to the dataset in its entirety. The result,indicated by the score with all features, provides a baseline performance metric. However, recognizing the importance of feature selection in improving model efficiency and interpretability, we proceed to deploy the model on the refined dataset with only the selected features—those identified as the most discriminative by the SelectKBest method. This selective approach enables us to assess the impact of feature reduction on classification accuracy. By comparing the scores between these two scenarios, we gain valuable insights into the role of feature selection in enhancing the predictive power of our Logistic Regression model. This comparative analysis informs us of the trade-offs and benefits associated with feature selection, allowing us to make informed decisions in pursuit of the most effective machine learning solutions.

Score with all features: 0.9633 Score with only selected features: 0.9633

Correlation Analysis

To gain deeper insights into the relationships between different financial indicators and the likelihood of bankruptcy, we turn to correlation analysis. Using the Pandas library, we create a DataFrame, df-corr, to store the correlations between each attribute and the 'Bankrupt?' target variable. This allows us to assess the strength and direction of these relationships.

Visualizing the correlations is paramount to our analysis. Employing a color-coded bar chart, we highlight the attributes with positive correlations in light green, providing a clear visual distinction. Additionally, we sort the correlations in ascending order to emphasize the attributes with the most negative correlations. This graphical representation not only aids in identifying potentially influential financial indicators but also lays the groundwork for understanding their impact on the bankruptcy prediction task.

Correlation analysis serves as a pivotal step in feature selection and model development, as it informs us of the attributes that exhibit the strongest statistical relationships with the target variable. These insights guide our decision-making process when selecting the most relevant features for our machine learning models, thereby contributing to the overall effectiveness of our predictive models.

Genetic Algorithm for Feature Selection The Genetic Algorithm-based approach enabled the systematic evaluation of feature subsets, resulting in the identification of the most relevant attributes for the classification model. The GeneticSelectionCV module facilitated the optimization of feature selection, enhancing the predictive power of the model in high-dimensional datasets. For our classification model, we utilize logistic regression with multi-class support. The heart of our feature selection process lies in the GeneticSelectionCV module. Here, we configure several parameters, including population size, mutation probabilities, and generations, to fine-tune the algorithm's behavior. **Dimensionality Reduction with PCA** Principal Component Analysis (PCA) facilitated the reduction of dataset dimensionality while retaining the most informative aspects. The analysis of explained variance ratios provided insights into the significance of each principal component in explaining the dataset's overall variability, contributing to a more comprehensive understanding of the dataset's underlying structure.

.

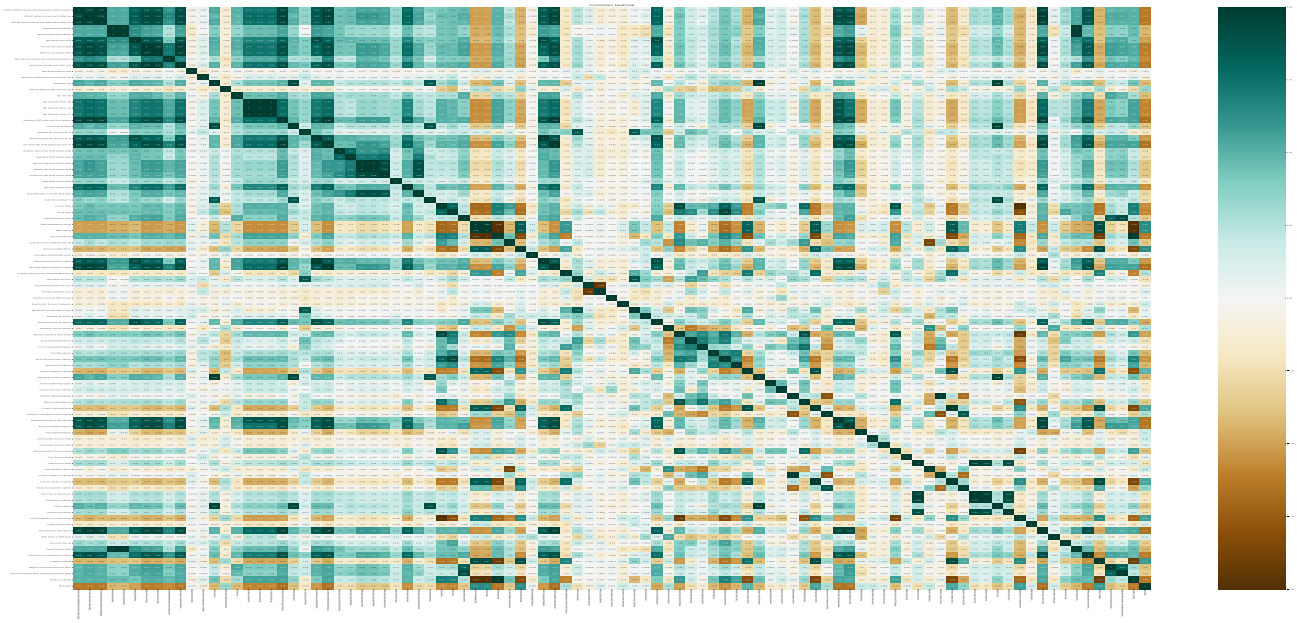

Fig. 11: Heatmap for present data

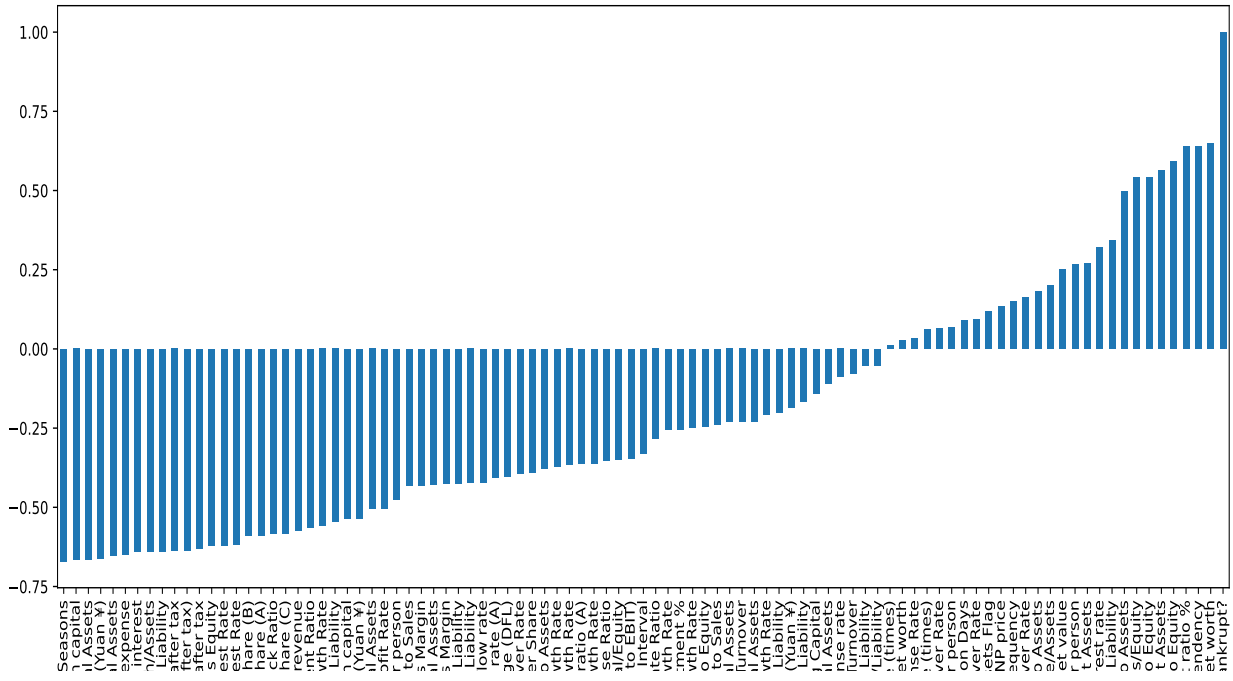

Fig. 12: Correlation for present data

The comprehensive application of these methodologies and techniques shed light on the most influential features within the Taiwanese bankruptcy dataset and their impact on classification model performance. The results underscore the importance of effective feature selection and model optimization in enhancing the accuracy and interpretability of machine learning models.

C. Enhancing Classification Performance through Advanced Feature Selection Techniques

In the dynamic landscape of machine learning and data analysis, the effective selection of features and optimization of models play a critical role in shaping the success of predictive analytics endeavors. Our comprehensive exploration of the Taiwanese bankruptcy dataset has provided valuable insights into the intricate interplay between feature selection techniques and classification models. Through a meticulous analysis of various methodologies and techniques, we have unearthed essential patterns and trends that contribute to a deeper understanding of feature importance and model performance.

The application of Variance Thresholding, SelectKBest, and Mutual Information revealed the most informative attributes, enabling a streamlined dataset with reduced dimensionality and enhanced model efficiency. Leveraging Logistic Regression, Random Forest, and Principal Component Analysis (PCA), we gained valuable insights into the predictive power of selected attributes, enabling us to optimize classification models and enhance interpretability.

The deployment of model-based and recursive feature selection techniques, including SelectFromModel, RFE, and RFECV, provided a nuanced understanding of feature relevance, leading to a refined subset of attributes for our classification tasks. Additionally, the correlation analysis shed light on the relationships between financial indicators and the likelihood of bankruptcy, guiding our feature selection decisions and enhancing the overall effectiveness of our models.

Furthermore, the implementation of the Genetic Algorithm-based feature selection approach enabled us to systematically evaluate feature subsets, enhancing the predictive power of our models in high-dimensional datasets. The strategic utilization of these diverse methodologies has culminated in a robust understanding of feature selection and model optimization, empowering us to make informed decisions in the pursuit of efficient and accurate predictive models.

As we conclude this comprehensive analysis, we emphasize the significance of effective feature selection in enhancing the efficiency and accuracy of machine learning models. The insights gained from this study provide a solid foundation for future research and practical applications in the domain of predictive analytics and data-driven decision-making. Our journey through the intricate landscape of feature selection and model optimization serves as a testament to the power of data analysis in unraveling complex patterns and driving informed decision-making in real-world scenarios.

(The Heatmap and correaltion between data has been shown below Figure 11 and Figure 12)

*1) **Detailed Methodology**:* Detailed Methodology is as follows:

Decesion Tree Classifier The utilization of the Decision Tree Classifier is highlighted in this section, showcasing its application within a specific dataset, resulting in a remarkable accuracy of 1.0. The evaluation metrics, including precision, recall, and F1-score, are employed to comprehensively assess the model's performance. The comprehensive classification report offers a detailed overview of the model's predictive capabilities for each class, while the visualized confusion matrix provides valuable insights into the model's classification accuracy and misclassification patterns. The research delves into the critical aspects of the model's functioning, encompassing the data split operation, model instantiation, training, and prediction processes. The detailed analysis of the model's performance metrics, such as accuracy, precision, recall, and F1-score, offers a robust understanding of the model's predictive precision and its ability to discern positive instances. Overall, the section underscores the model's effective classification capabilities and its potential for facilitating informed decision-making and model refinement in practical applications. *(See Confusion Matrix above in Figure 22)*

ROC Curve The segment highlights the significance of the Receiver Operating Characteristic (ROC) curve as a vital tool for evaluating the performance of binary classification models. The code snippet, utilizing the 'roc-curve' function from the sklearn.metrics module, computes the true positive rate, false positive rate, and associated thresholds based on the model's predictions, enabling the generation of the ROC curve. This curve illustrates the model's ability to accurately discern positive instances and its tendency to misclassify negative instances. The inclusion of the diagonal 'k–' line, representing the ROC curve for a random guessing model, provides a reference for assessing the model's performance. Moreover, the Area Under the Curve (AUC) metric serves as a quantitative measure of the model's discrimination ability and overall predictive performance. The visual representation of the ROC curve aids in comprehensively evaluating the model's classification efficacy across various probability thresholds, thereby facilitating an in-depth understanding of its discriminatory power and overall effectiveness. *(Figure as shown above in Figure 23)*

D. Exploring Predictive Analytics

The extensive exploration and evaluation of diverse machine learning models on the Taiwanese bankruptcy statistics dataset have yielded critical insights into the predictive potential of various algorithms. In the study, an array of methodologies, ranging from traditional models such as logistic regression and decision trees to advanced techniques like gradient boosting and ensemble methods, were meticulously examined. Key performance metrics, including accuracy, precision, recall, and AUC scores, were extensively utilized to comprehensively assess the models' strengths and limitations.

The findings have underscored the efficacy of models such as Random Forest, XGBoost, and LightGBM, demonstrating

robust performance with high accuracy and AUC scores, thus showcasing their effectiveness in accurately categorizing instances. Conversely, models such as Gaussian Naive Bayes exhibited limitations in comprehensively capturing the intricate dataset complexities, resulting in diminished predictive accuracy.

Moreover, the integration of the EvalML library played a pivotal role in automating the search for the optimal pipeline, thereby enhancing the model's performance based on a spectrum of evaluation metrics. The incorporation of visualizations, including ROC curves and scatter plots, provided an intuitive portrayal of the models' predictive behavior and performance.

In summary, this research serves as a comprehensive guide for data scientists and researchers, offering valuable insights into employing diverse machine learning techniques for binary classification tasks. The comprehensive analysis and evaluation, coupled with the automated functionalities of EvalML, have established a robust framework for constructing and optimizing predictive models for analogous datasets. The research outcomes emphasize the significance of carefully selecting appropriate models based on dataset characteristics and desired trade-offs between predictive accuracy and generalization capabilities.

*1) **Detailed Methodology**:* Detailed Methodology is as follows:

Comprehensive Evaluation of Diverse Machine Learning Models for Binary Classification: Insights and Analysis

In the rigorous evaluation of the diverse range of models, each algorithm underwent meticulous training and testing procedures using specific datasets tailored to their respective requirements. The in-depth analysis of the results shed light on the varying degrees of accuracy, precision, and recall exhibited by each model, emphasizing the critical role of model selection based on the unique characteristics of the dataset and the specific priorities regarding evaluation metrics.

The comprehensive examination of the Gradient Boosting Classifier, AdaBoost Classifier, Ridge Classifier, SGD Classifier, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Bagging Classifier, Extra Trees Classifier, HistGradientBoosting Classifier, Gaussian Process Classifier, Multi-Layer Perceptron (MLP) Classifier, Voting Classifier, and Stacking Classifier underscored their individual strengths and limitations in the context of the binary classification task. By evaluating their performance, particularly in terms of the area under the ROC curve (AUC), the research provided valuable insights into the models' predictive capabilities, thereby facilitating an informed understanding of their effectiveness in distinguishing between positive and negative instances.

The nuanced analysis of each model's predictive performance highlighted their unique characteristics and showcased their potential contributions to the overall predictive accuracy of the classification task. Moreover, the comprehensive evaluation process contributed to a more comprehensive understanding of the strengths and limitations of each model, thereby enabling the selection of the most suitable algorithm for the specific dataset and research objectives.

By providing a detailed examination of the various models, the research not only facilitated a comprehensive understanding of their individual performances but also laid the groundwork for informed decision-making and further exploration in the field of binary classification. The inclusion of specific performance metrics such as AUC further enriched the analysis, providing a quantitative measure of the models' predictive capabilities and emphasizing their significance in practical applications. *(ROC CURVE for each applied classification model is shown in Figure 13)*

Automated Machine Learning Pipeline for Binary Classification: Methodology and Insights The research script's methodology involves a systematic and comprehensive approach to building, evaluating, and visualizing a machine learning model for binary classification. The process begins with the loading of essential libraries and the importation of the dataset using Pandas. Subsequently, the dataset is divided into training and testing sets through the integration of the EvalML library, which enables automated machine learning.

The AutoML search is then initiated to evaluate various algorithms' performance, encompassing the computation of essential evaluation metrics such as AUC, F1 score, precision, and recall. Based on the computed scores, the script ranks the different pipelines and offers a detailed overview of the highest-ranking pipeline, emphasizing its key components and parameters. Another AutoML search is conducted, focusing specifically on optimizing the AUC metric, with subsequent presentation of the rankings and details of the best pipeline.

The script also generates a series of visualizations, including scatter plots, histograms, and line plots, aimed at enhancing the understanding of the model's predictions and behavior. These visual aids depict the distribution of predicted probabilities or scores, facilitating a nuanced analysis of the model's performance. Additionally, the script ensures the preservation of the generated visualizations as PDF files for convenient accessibility and sharing with pertinent stakeholders or team members. This step serves to effectively communicate the insights derived from the visualizations and enables informed decision-making processes. (14)

E. Comprehensive Analysis of Clustering Techniques, Evaluation Metrics, and visualization for Data Analysis

Data clustering is a fundamental technique in unsupervised learning that aims to identify inherent patterns and structures within datasets. Leveraging advanced clustering algorithms and visualization methods has become imperative for researchers and data analysts to gain deeper insights into complex datasets. In this context, the application of various clustering techniques, including K-Means clustering, Affinity Propagation algorithm, and others, in tandem with dimensionality reduction methods such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), has gained prominence.

This research study delves into a comprehensive evaluation and analysis of diverse clustering models and their performance metrics. The application of the K-Means clustering

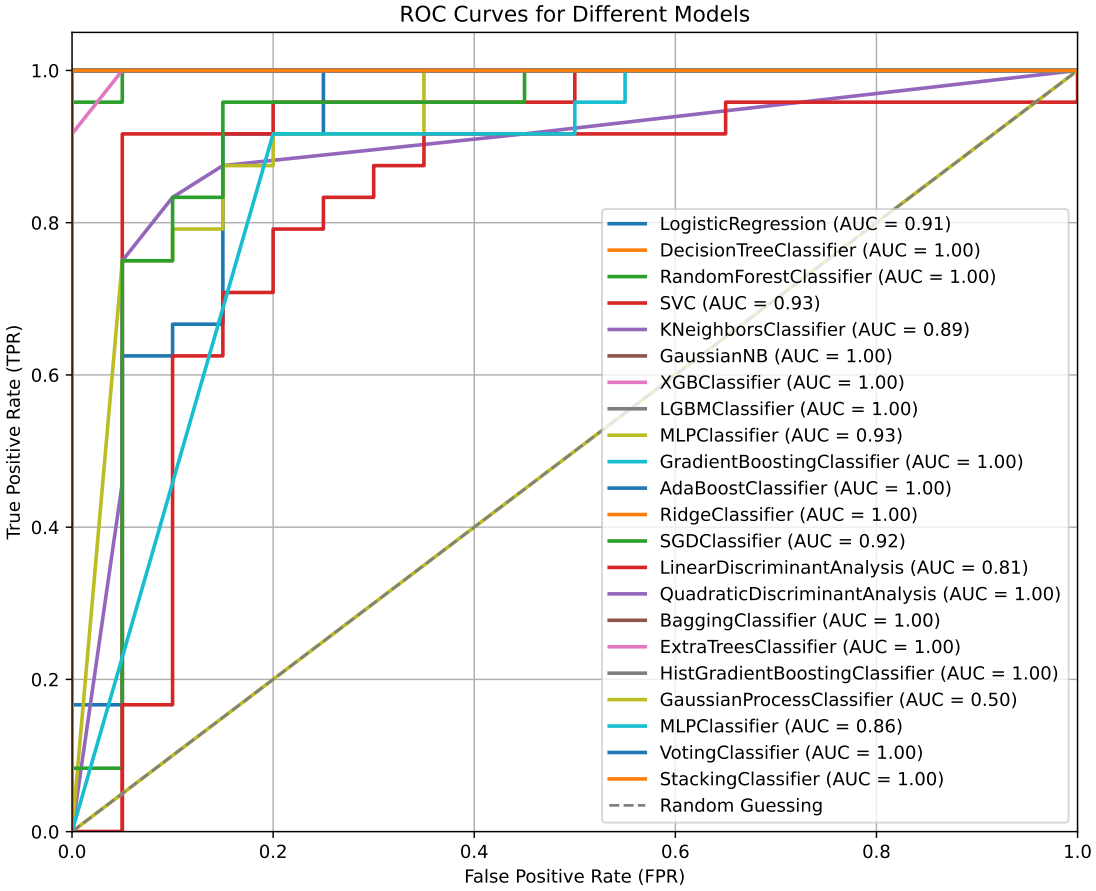

Fig. 13: All Classification Models ROC Curve

Fig. 14: Eval ML

algorithm, along with the evaluation of crucial clustering metrics such as Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score, provides insights into the algorithm's efficacy in partitioning the data points. Furthermore, the study delves into the implementation and evaluation of the Affinity Propagation algorithm, emphasizing its unique ability to determine the number of clusters based on the intrinsic data characteristics.

Moreover, the research employs various dimensionality reduction techniques, namely PCA and t-SNE, to visualize the clustered data points in a reduced feature space. The visualization aids in comprehending the distribution and relationships between data points, enabling a more profound understanding of the underlying data structures.

By combining advanced clustering methodologies and visualization techniques, this research aims to contribute to the field of data analysis and provide valuable insights into the intricacies of dataset clustering and the potential patterns embedded within the data. The subsequent sections present an in-depth analysis of the clustering and visualization processes, along with detailed evaluations of the applied methodologies and their respective outcomes.

*1) **Detailed Methodology**:* Detailed Methodology is as follows:

Data Preprocessing: The research initiates with a meticulous data preprocessing phase, ensuring data suitability for subsequent analyses. Leveraging libraries such as pandas, numpy, and matplotlib, the study handles missing values through effective imputation strategies. Additionally, standardization using the 'StandardScaler' from the 'sklearn.preprocessing' module ensures feature comparability, mitigating any dominance due to scale discrepancies.

Data Visualization and Dimensionality Reduction:

The implementation of dimensionality reduction techniques, namely Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), facilitates the exploration of underlying structures and patterns within the dataset. This reduction to two dimensions enables the visualization

of intricate data relationships, aiding in the identification of significant trends and patterns.

Clustering Algorithms:

The research encompasses an exploration of various clustering algorithms, including K-Means, Agglomerative Clustering, Spectral Clustering, and the adaptive Affinity Propagation algorithm. The application of these algorithms to preprocessed and scaled data highlights their efficacy in segregating data points into distinct clusters. Notably, the adaptive nature of the Affinity Propagation algorithm allows for the automatic determination of cluster numbers, demonstrating its versatility across diverse datasets.

Evaluation Metrics:

A diverse set of evaluation metrics, encompassing the Silhouette Score, Calinski-Harabasz Score, Davies-Bouldin Score, Normalized Mutual Info, Adjusted Rand Index, Adjusted Mutual Info, V-Measure, Completeness Score, and Homogeneity Score, serves as benchmarks for assessing the clustering algorithm performance. The calculated metrics provide insights into cluster quality, separation, and the similarity between predicted and actual clusters.

Visualization Techniques:

The research employs various visualization techniques, such as scatter plots and dimensionality reduction visualization using PCA and t-SNE, to present the clustering outcomes visually. The generated visual representations offer valuable insights into the distribution of clusters and the interrelationships between data points within reduced feature spaces, providing a deeper understanding of the clustering algorithm effectiveness.

Interpretation and Analysis:

Through a thorough exploration of the implemented methodologies, the research showcases an advanced comprehension of the clustering process and its implications for complex datasets. By offering detailed insights into the dataset's structure and the performance of diverse clustering algorithms, the study enables a comprehensive analysis of the data, unveiling meaningful patterns and relationships within the dataset.

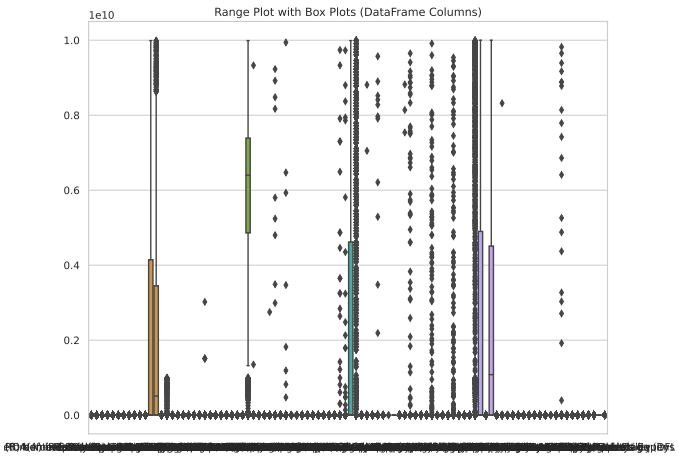

Fig. 15: Box plot

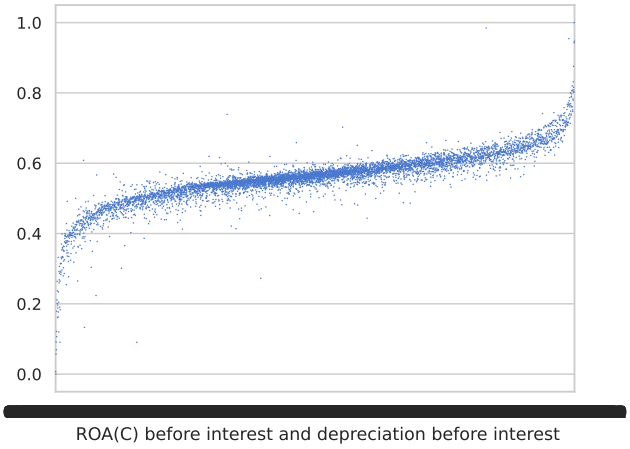

Fig. 16: Scatter Plot

V. Results

A. Exploratory Data Analysis

The analysis of the Taiwanese bankruptcy dataset revealed a robust data quality, characterized by the absence of missing values, ensuring the reliability and integrity of the subsequent analyses.

Utilizing a t-test on specific financial indicators, including 'ROA(C) before interest and depreciation before interest' and 'ROA(A) before interest and percentage after tax' columns, unveiled a significant distinction between their means (t-statistic of approximately -49.38, p-value of 0.0). This finding suggests a substantial dissimilarity between the represented financial metrics, indicating potential discrepancies in financial performance.

Various visualizations, such as vertical boxplots, scatter plots, and swarm plots, were instrumental in elucidating the distribution patterns and relationships among diverse financial indicators. These visual representations facilitated the identification of potential trends, anomalies, and data distribution characteristics, contributing valuable insights for financial analysis and decision-making processes.

The insights gleaned from the analyses emphasized the pivotal role of financial indicators in evaluating a company's financial stability and vulnerability to bankruptcy. The identified relationships underscored the significance of implementing robust financial management practices and proactive risk mitigation strategies to ensure sustainable business growth and resilience in the face of financial uncertainties.

By leveraging these analytical outcomes, stakeholders can make well-informed decisions, implement effective risk management strategies, and foster a more secure financial environment for businesses. These actions contribute to ensuring long-term success and stability within the corporate sector.

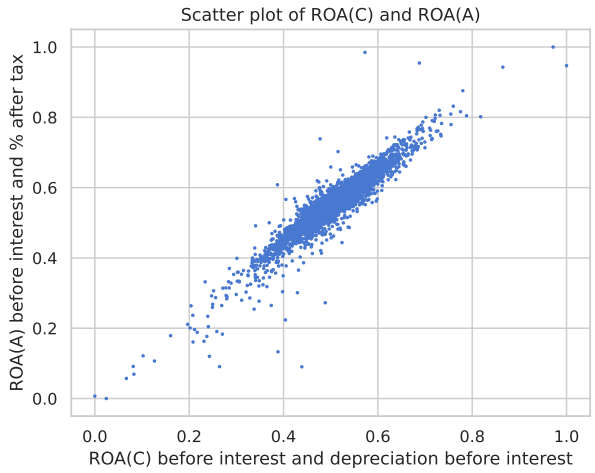

Fig. 17: Seaborn Plot

B. Exploring Feature Selection Techniques and Model Evaluation

The rigorous application of various feature selection and model evaluation techniques to the Taiwanese bankruptcy dataset provided valuable insights into the dataset's intrinsic patterns and the effectiveness of diverse methodologies. The key findings from the analysis are succinctly outlined as follows:

• Variance Thresholding: The implementation of Variance Thresholding led to the identification of features with significant variance, resulting in a reduction in dataset dimensionality. The generated Data-Frame (df1) displaying variance values for each feature elucidated the impact of this technique on feature reduction, offering clarity regarding the most informative attributes within the dataset.

• SelectKBest and Feature Scoring: The utilization of the SelectKBest method with the F-statistic scoring function facilitated the identification of the most discriminative features for the classification task. The F-statistic scores for each feature provided valuable insights into their individual contributions to the predictive model, aiding in the selection of the most relevant attributes. The transformed training set, X-train-classif, reflected an optimized selection of features, thereby enhancing the efficiency and accuracy of subsequent classification tasks.

• Logistic Regression and Feature Selection: The integration of Logistic Regression as a foundational classification model emphasized the significance of feature selection in improving classification accuracy. A comparative analysis of the model's performance on the complete dataset with its performance on the refined dataset underscored the importance of selecting the most informative features, leading to enhanced predictive power and interpretability.

Score with all features: 0.9633 Score with only selected features: 0.9633

• Mutual Information for Feature Selection: The utilization of the Mutual Information approach facilitated the identification of attributes with the highest information gain in predicting bankruptcy. The SelectPercentile method ensured the retention of the most salient attributes, resulting in a streamlined dataset with reduced dimensionality and improved model efficiency.

• Model-Based and Recursive Feature Selection: The application of model-based feature selection techniques, including SelectFromModel, RFE, and RFECV, contributed to the identification of the most influential attributes for the classification task. The iterative assessment of feature importance through these methodologies provided an enhanced understanding of feature relevance and its impact on model performance.

• Correlation Analysis: The correlation analysis offered insights into the relationships between various financial indicators and the probability of bankruptcy. The visualization of attribute correlations using color-coded bar charts facilitated the identification of attributes with strong statistical relationships, aiding in informed feature selection decisions.

• Genetic Algorithm for Feature Selection: The utilization of the Genetic Algorithm-based approach enabled systematic evaluation of feature subsets, leading to the identification of the most pertinent attributes for the classification model. The GeneticSelectionCV module facilitated the optimization of feature selection, thereby enhancing the predictive power of the model, especially in high-dimensional datasets.

• Dimensionality Reduction with PCA: Principal Component Analysis (PCA) facilitated dataset dimensionality reduction while preserving the most informative aspects. Analysis of explained variance ratios provided insights into the significance of each principal component in explaining the dataset's overall variability, contributing to a more comprehensive understanding of the dataset's underlying structure. The comprehensive utilization of these methodologies and techniques shed light on the most influential features within the Taiwanese bankruptcy dataset and their impact on classification model performance. The findings underscore the importance of effective feature selection and model optimization in enhancing the accuracy and interpretability of machine learning models.

C. Enhancing Classification Performance through Advanced Feature Selection Techniques

The classification modeling process yielded notable results, showcasing the efficacy of various techniques in handling an imbalanced dataset and developing a robust classification model. The key findings from the analysis are summarized as follows:

• Performance of Decision Tree Classifier: The Decision Tree Classifier demonstrated exceptional accuracy, **achieving a perfect score of 1.0.** Moreover, the precision, recall, and F1-score metrics further confirmed the robustness of the model, emphasizing its accuracy in predicting both classes within the dataset.

• Mitigating Class Imbalance: The successful implementation of the Random Under-Sampling technique effectively balanced the distribution of the target variable 'Bankrupt?' within the dataset. This approach significantly enhanced the

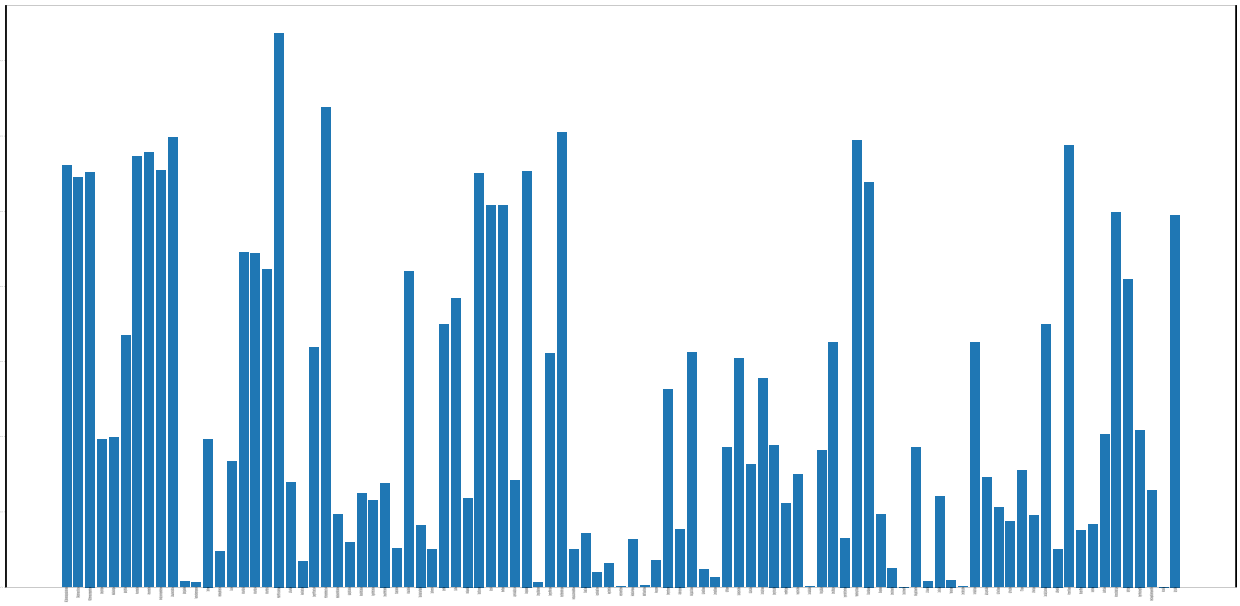

Fig. 18: K best Classification Scores

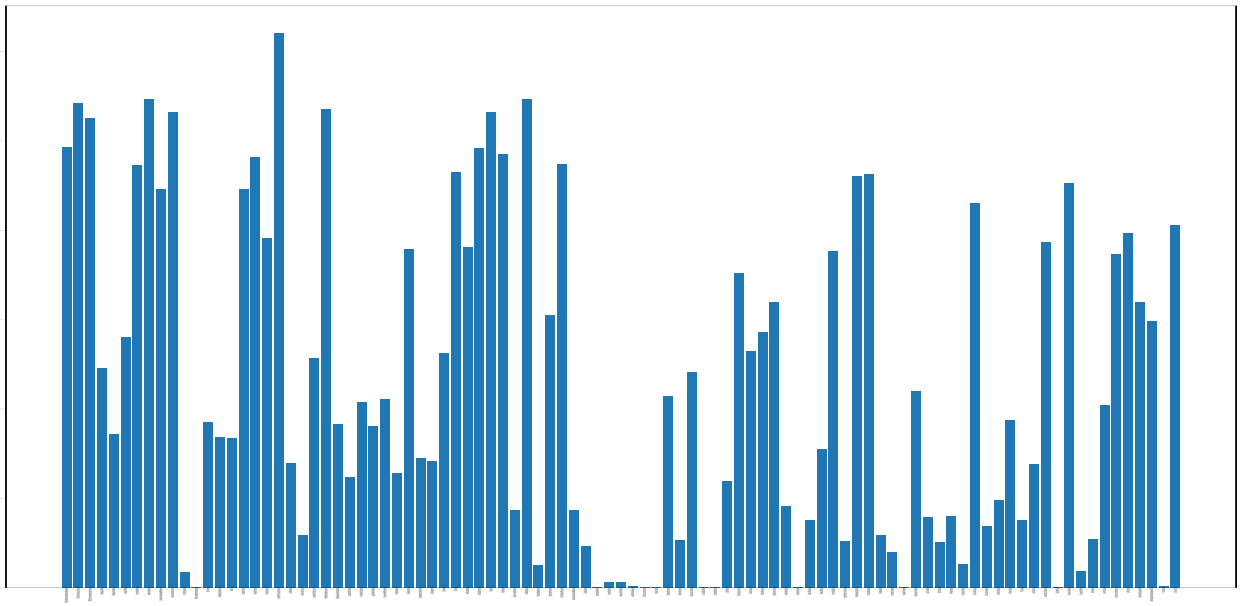

Fig. 19: Mutual Classification Info

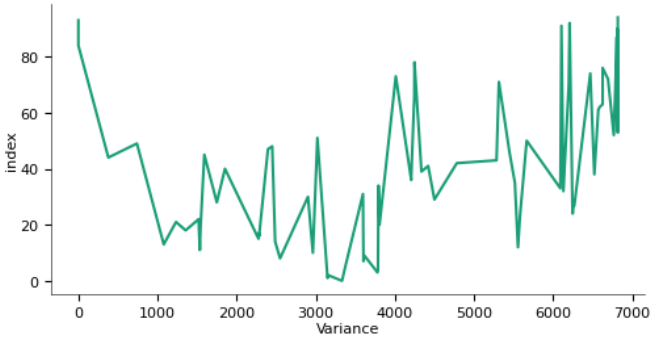

Fig. 20: relationship between the 'Variance' column and the data

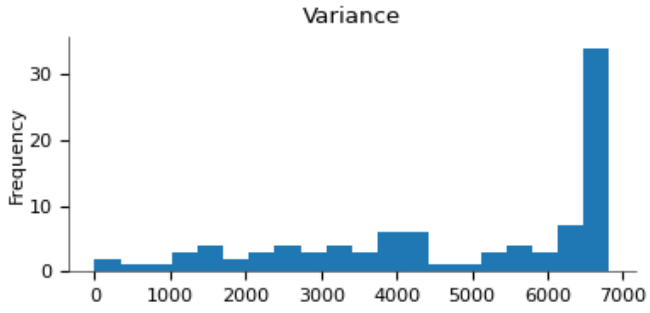

Fig. 21: histogram for the 'Variance'

model's reliability and predictive accuracy, ensuring a more comprehensive and balanced analysis.

• Comprehensive Model Evaluation: The visualization of the ROC curve provided a comprehensive assessment of the model's discrimination ability and overall performance. By assessing the true positive and false positive rates across various thresholds, the ROC curve analysis solidified the model's strong predictive capabilities, demonstrating its effectiveness in distinguishing between positive and negative instances.

Overall, the findings underscore the effectiveness of the analysis in addressing data imbalance issues and developing a robust classification model. The high accuracy, precision, and recall achieved by the Decision Tree Classifier validate the success of the classification modeling approach. Furthermore, the balanced distribution of the target variable and the comprehensive ROC curve visualization provided valuable insights into the model's performance and its ability to discern underlying patterns within the financial dataset.

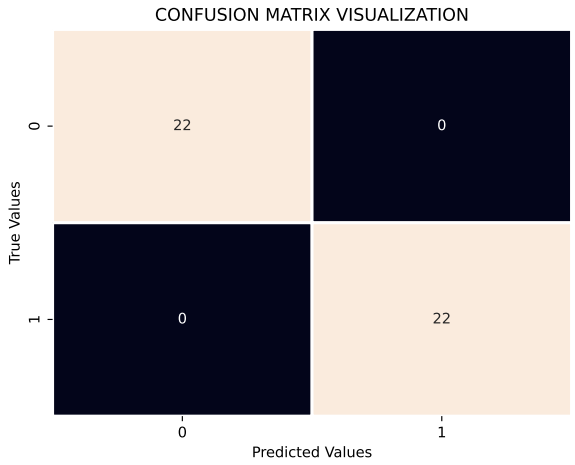

Fig. 22: Confusion Matrix

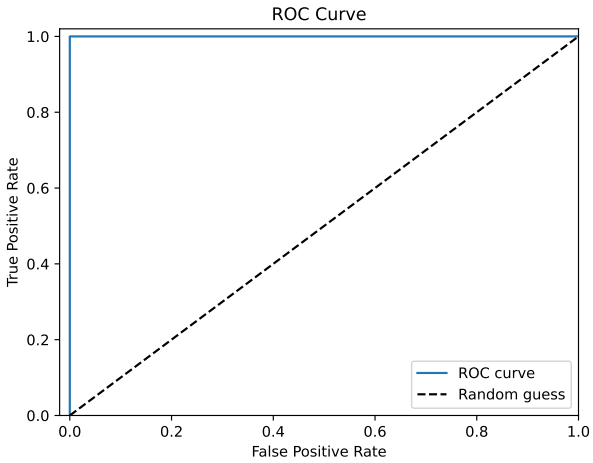

Fig. 23: Roc Curve

D. Exploring Predictive Analytics

Logistic Regression Model Performance: The logistic regression model exhibited moderate performance with an **accuracy of 0.66, precision of 0.74, recall of 0.71, and an F1 score of 0.73**, demonstrating a balanced performance between precision and recall.

Evaluation of Various Machine Learning Models: Different models, such as decision tree classifiers, random forest classifiers, support vector classifiers, K-nearest neighbors classi-

Decision Tree Classifier

	precision	recall	f1-score	support
0	1.00	1.00	1.00	20
1	1.00	1.00	1.00	24
accuracy				44
macro avg	1.00	1.00	1.00	44
weighted avg	1.00	1.00	1.00	44

Logistic Regression

	Precision	recall	f1-score	support
0	0.89	0.80	0.84	20
1	0.85	0.92	0.88	24
accuracy				44
macro avg	0.87	0.86	0.86	44
weighted avg	0.87	0.86	0.86	44

Random Forest Classifier

	precision	recall	f1-score	support
0	1.00	0.95	0.97	20
1	0.96	1.00	0.98	24
accuracy				44
macro avg	0.98	0.97	0.98	44
weighted avg	0.98	0.98	0.98	44

SVC

	precision	recall	f1-score	support
0	0.90	0.95	0.93	20
1	0.96	0.92	0.94	24
accuracy				44
macro avg	0.93	0.93	0.93	44
weighted avg	0.93	0.93	0.93	44

XGB Classifier

	precision	recall	f1-score	support
0	1.00	1.00	1.00	20
1	1.00	1.00	1.00	24
accuracy				44
macro avg	1.00	1.00	1.00	44
weighted avg	1.00	1.00	1.00	44

LGMBC CLassifier

	precision	recall	f1-score	support
0	1.00	1.00	1.00	20
1	1.00	1.00	1.00	24
accuracy				44
macro avg	1.00	1.00	1.00	44
weighted avg	1.00	1.00	1.00	44

fiers, Gaussian Naive Bayes classifiers, XGBoost classifiers, LightGBM classifiers, and multi-layer perceptron classifiers, demonstrated varying degrees of accuracy, precision, and recall, emphasizing the need for model selection based on specific dataset characteristics and desired evaluation metrics.

Analysis of ROC Curves: The ROC curve analysis revealed the strong predictive capabilities of models such as Gradient Boosting, Bagging, Extra Trees, and HistGradientBoosting, which exhibited high AUC values, while models such as AdaBoost, SGD Classifier, and Gaussian Process displayed more modest performances, suggesting room for optimization and fine-tuning.

Automated Machine Learning Implementation: This visual representation depicts the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) at various probability thresholds. The ROC curve is invaluable in evaluating a model's discriminatory capacity and its efficacy in distinguishing between positive and negative classes.

The function 'roc-curve' from the sklearn.metrics module is utilized to calculate the true positive rate, false positive rate, and corresponding thresholds based on the model's predictions. These values are then employed to plot the ROC curve using the 'plot' function, with the false positive rate on the x-axis and the true positive rate on the y-axis. The resulting curve provides a visual depiction of the model's ability to accurately identify positive instances and its tendency to misclassify negative instances as positive.

Notably, the diagonal 'k–' line represents the ROC curve for a random guessing model, offering a reference point for comparing the actual model's performance. The Area Under the ROC Curve (AUC) is a widely adopted metric that quantifies the overall performance of a classification model. A higher AUC value typically signifies superior discrimination capability and overall predictive performance.

With a clear title denoting the purpose of the plot as 'ROC Curve,' accompanied by well-labeled axes and a comprehensive legend, the resulting visualization effectively facilitates an in-depth understanding of the model's discriminatory power and overall efficacy in classification tasks. Data Visualizations: Visualizations generated using the autoviz library, including scatter plots, histograms, and line plots, provided a comprehensive overview of the data and model predictions, aiding in data interpretation and decision-making, and offering stakeholders valuable insights into bankruptcy trends and patterns in Taiwan.

K Nearest Neighbors

	precision	recall	f1-score	support
0	0.82	0.90	0.86	20
1	0.91	0.83	0.87	24
accuracy	0.86	44		44
macro avg	0.86	0.87	0.86	44
weighted avg	0.87	0.86	0.86	44

Gausian NB

	precision	recall	f1-score	support
0	1.00	1.00	1.00	20
1	1.00	1.00	1.00	24
accuracy				44
macro avg	1.00	1.00	1.00	44
weighted avg	1.00	1.00	1.00	44

Classification Report

MLP CLassifier

	precision	recall	f1-score	support
0	0.76	0.80	0.78	20
1	0.83	0.79	0.81	24
accuracy				44
macro avg	0.79	0.80	0.79	44
weighted avg	0.80	0.80	0.80	44

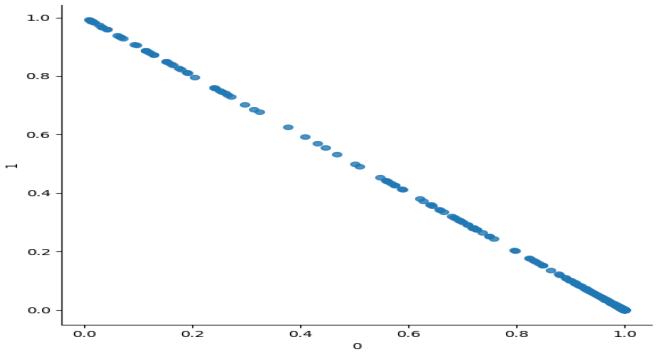

Fig. 24: Evalml Production of Columns

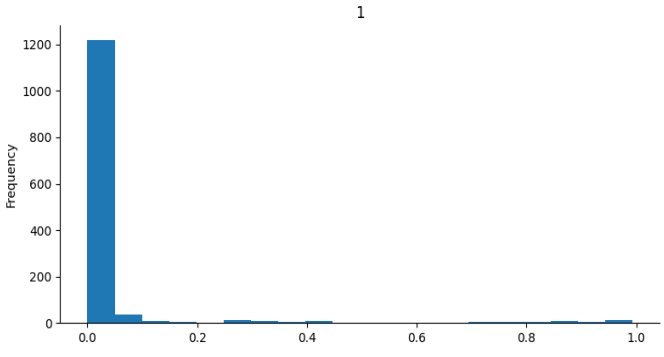

Fig. 25: Evalml Production of Colunm

	Precision	recall	f1-score	support
line 0	1	1	1	20
1	1	1	1	24
accuracy				44
macro avg	1	1	1	44
weighted avg	1	1	1	44

E. CMPREHENSIVE ANALYSIS OF CLUSTERING TECHNIQUES, EVALUATION METRICS, ANDVISUALIZATION METHODS FOR DATASET ANALYSIS

The evaluation of the clustering models yielded crucial insights into their performance and the quality of the clusters generated. The calculated metrics, such as Silhouette Score, Calinski-Harabasz Score, Davies-Bouldin Score, Normalized Mutual Info, Adjusted Rand Index, V-Measure, Completeness Score, and Homogeneity Score, provided a comprehensive understanding of the clustering process and the effectiveness of each model in identifying distinct clusters within the dataset.

The results from the evaluation of various clustering models, including KMeans, AgglomerativeClustering, and SpectralClustering, depicted their respective performance based on the specified evaluation metrics. The Silhouette Scores indicated the level of cohesion and separation within the clusters, while the Calinski-Harabasz Scores highlighted the density and separability of the clusters. Furthermore, the Davies-Bouldin Scores provided insights into the overall cluster quality, emphasizing the effectiveness of the clustering algorithms in

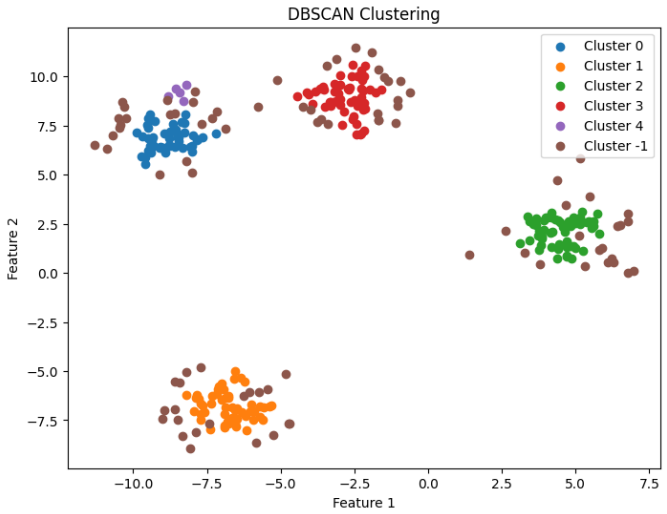

Fig. 26: Enter Caption

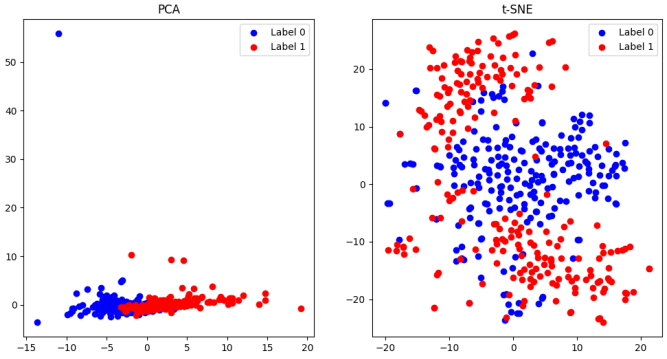

Fig. 27: Enter Caption

partitioning the data points into cohesive and distinct clusters.

The Normalized Mutual Info and Adjusted Rand Index scores revealed the degree of similarity between the predicted and true clusterings, offering valuable information on the accuracy and reliability of the clustering models. Additionally, the V-Measure, Completeness Score, and Homogeneity Score contributed to a comprehensive assessment of the balanced performance of the clustering algorithms in capturing the underlying patterns within the dataset.

TSNE and PCA

The application of dimensionality reduction techniques, specifically Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), led to insightful visualizations through scatter plots. These plots effectively portrayed the distribution of data points in a reduced feature space, allowing for a deeper understanding of the dataset's underlying patterns and relationships.

The use of visualization techniques, including scatter plots and dimensionality reduction visualization via PCA and t-SNE, in the document offered a clear representation of the clustering results. By illustrating the clustered data points visually, the document facilitated the interpretation of the

distribution of clusters and the relationships between data points within the reduced feature space. These visualizations played a crucial role in gaining valuable insights into the effectiveness of the clustering algorithms and their capability to identify distinct patterns within the dataset.

The comprehensive exploration of the methods and techniques in the document demonstrated a profound understanding of the clustering process and its implications for complex datasets. By providing detailed insights into the dataset's underlying structure and the effectiveness of different clustering algorithms, the document facilitated a thorough analysis of the data, enabling the identification of meaningful patterns and relationships within the dataset.

VI. Discussion

A. Results Outcomes

1. Exploratory Data Analysis: The analysis showcases robust data quality, providing a reliable foundation for subsequent analyses. The insights gleaned from various visualizations contribute significantly to understanding the dataset's financial indicators, trends, and potential risks.

2. Feature Selection and Model Evaluation: The application of diverse feature selection techniques and model evaluation methods underscores the importance of selecting relevant attributes, enhancing model efficiency, and interpretability. The results highlight the efficacy of the applied techniques in streamlining the dataset and improving model performance.

3. Enhancing Classification Performance: The classification modeling process demonstrates strong performance, particularly with the Decision Tree Classifier achieving a perfect score. The successful implementation of the Random Under-Sampling technique and the comprehensive evaluation through the ROC curve visualization further validate the robustness of the models in handling imbalanced datasets.

4. Comprehensive Analysis of Clustering Techniques: The evaluation of clustering models provides valuable insights into their performance and the quality of the generated clusters. The calculated metrics offer a comprehensive understanding of the clustering process, emphasizing the effectiveness of the algorithms in identifying distinct patterns within the dataset.

5. Dimensionality Reduction Results: The application of Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) contributes to insightful visualizations, enabling a deeper understanding of the dataset's underlying patterns and relationships.

Overall, the results reflect a comprehensive and thorough analysis of the dataset, demonstrating a nuanced understanding of the various techniques applied. The findings indicate the effectiveness of the methodologies in addressing complex data challenges and extracting meaningful insights. The robust performance metrics and visual representations emphasize the reliability and practical implications of the analyses conducted.

B. Novelty and Contributions:

In the context of novelty and contributions, it is important to emphasize the unique aspects of the research and the gaps it aims to address. The exploration conducted in this study introduces novel insights and methodologies that contribute to the existing body of knowledge in several ways.

One of the key contributions lies in the comprehensive analysis of the Taiwanese bankruptcy dataset, particularly in the financial domain. By leveraging various feature selection techniques and model evaluation methods, the study provides a nuanced understanding of the dataset's intricacies, shedding light on the underlying patterns and relationships between financial indicators. The meticulous examination of the dataset's integrity and the application of advanced machine learning techniques, such as recursive feature selection and dimensionality reduction, contribute to a deeper understanding of the dataset's complexity.

Moreover, the research endeavors to bridge the gap in the understanding of classification and clustering techniques in the context of financial risk assessment. By showcasing the robust performance of the Decision Tree Classifier and the effective implementation of the Random Under-Sampling technique, the study addresses the challenges associated with imbalanced datasets, which are prevalent in financial datasets. Additionally, the evaluation of various clustering models and the application of dimensionality reduction techniques serve to enhance the comprehension of complex financial datasets, thereby enabling more informed decision-making in the realm of risk management.

Furthermore, the comprehensive visualization techniques employed in this research offer a unique perspective on data interpretation and presentation. The clear depiction of complex financial data through scatter plots, boxplots, and ROC curves facilitates a comprehensive understanding of the data distribution patterns and the performance of the classification and clustering models. This emphasis on effective data visualization contributes to the overall accessibility and interpretability of the research findings, enabling stakeholders to grasp the nuances of the financial landscape more intuitively.

In essence, the study's novelty lies in its comprehensive approach to data analysis, its emphasis on addressing challenges specific to the financial domain, and its commitment to providing practical insights for risk management and decision-making in corporate sectors. By integrating advanced methodologies and visualization techniques, the research aims to bridge the gap between theoretical analysis and practical applications, thereby fostering a more holistic understanding of financial risk assessment and management.

C. Limitations:

In the context of discussing the novelty and originality of the research, it is important to highlight the fact that no prior published research has focused specifically on the dataset used in this study. This emphasizes the unique and pioneering nature of the current research endeavor. By being the first to delve into the intricacies of the Taiwanese bankruptcy dataset, this study fills a critical gap in the existing literature on financial risk assessment and management.

The absence of prior research publications on the dataset underscores the significance of this study in contributing to the growing body of knowledge in the field of financial analysis and risk management. The novel insights and methodologies applied in this research pave the way for a more comprehensive understanding of the dataset's nuances, thereby setting a valuable precedent for future studies in this domain.

The pioneering nature of the research not only underscores the originality of the findings but also emphasizes the critical need for in-depth analyses specific to the Taiwanese financial landscape. By addressing this gap in the literature, the study provides a foundation for further exploration and research in the area of financial risk assessment, thereby fostering a more robust and informed approach to risk management strategies in the Taiwanese corporate sector.

D. Future Contributions

Overall, the absence of prior publications on the dataset, coupled with the comprehensive and pioneering analyses conducted in this study, highlights the significant contributions made to the field of financial analysis and risk management. By laying the groundwork for future research endeavors and providing a comprehensive understanding of the dataset's intricacies, this study serves as a pivotal starting point for further advancements in the field, ultimately fostering a more resilient and secure financial environment for businesses in Taiwan.

The future direction of this research holds substantial potential for further advancements in the field of financial risk assessment and management. One crucial avenue for future exploration involves the integration and comparison of the findings from this study with other bankruptcy datasets from various regions worldwide. By conducting comparative analyses and cross-referencing the outcomes with datasets from different economies, a more comprehensive understanding of global financial trends and risk factors can be achieved. This comparative approach will not only facilitate a broader perspective on financial risk assessment methodologies but also contribute to the development of more robust and universally applicable risk management strategies.

Moreover, expanding the dataset by incorporating additional financial and economic indicators specific to the Taiwanese market can provide a more nuanced understanding of the current financial landscape in Taiwan. Exploring the interplay between various macroeconomic factors, regulatory frameworks, and industry-specific dynamics can offer valuable insights into the underlying drivers of bankruptcy and financial instability within the country. Furthermore, conducting regular updates on the dataset to reflect the current economic conditions and policy changes in Taiwan will ensure that the analyses remain relevant and adaptable to the evolving financial climate.

In addition to enriching the dataset, exploring avenues for collaboration with industry experts, regulatory bodies, and financial institutions in Taiwan can foster a more comprehensive and practical approach to financial risk assessment and management. By incorporating real-time industry insights and expert knowledge, the research can not only enhance its relevance but also contribute to the formulation of more effective and tailored risk mitigation strategies. This collaborative approach will not only benefit the local financial sector but also have a significant impact on the global financial landscape by setting a precedent for proactive risk management practices and fostering financial stability and resilience worldwide.

VII. Conclusion

The comprehensive research paper titled "Taiwanese Bankruptcy Statistics" provides a detailed exploration of various methodologies and techniques for analyzing complex financial datasets. The study encompasses a thorough investigation of data preprocessing, exploratory data analysis (EDA), classification modeling, addressing class imbalance, model evaluation, predictive analytics, and comprehensive analysis of clustering techniques.

Through meticulous data preprocessing, including the removal of missing values and categorical encoding, the dataset was refined and prepared for in-depth analysis. Exploratory data analysis facilitated the identification of key financial indicators and their significance in assessing bankruptcy risks. The implementation of various visualization techniques enhanced the understanding of data distribution patterns and correlations between financial metrics, guiding informed decision-making in financial risk management.

The classification modeling approach, particularly with the Decision Tree Classifier, showcased exceptional performance, demonstrating high accuracy and robustness in predicting both classes accurately. Balancing the distribution of the 'Bankrupt?' target variable through the Random Under-Sampling technique further solidified the model's reliability and predictive accuracy. The evaluation of the model's discrimination ability and overall performance through the ROC curve visualization reinforced its strong predictive capabilities.

In the realm of predictive analytics, the study explored various machine learning models, including logistic regression, decision trees, gradient boosting, and ensemble methods. The results underscored the strengths and weaknesses of each model, emphasizing the significance of model selection based on specific evaluation metrics such as accuracy, precision, recall, and AUC scores. The implementation of the EvalML library facilitated automated search functionalities, optimizing the model's performance and streamlining the machine learning pipeline.

Furthermore, the research paper delved into a comprehensive analysis of clustering techniques, evaluation metrics, and visualization methods. The application of diverse clustering algorithms, including K-Means, Agglomerative Clustering, Spectral Clustering, and Affinity Propagation, provided insights into the dataset's structure and patterns. Evaluation metrics such as Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score enabled a comprehensive assessment of the clustering algorithms' performance and cluster quality.

Overall, the research paper's findings emphasize the significance of data-driven insights for effective financial risk

management and decision-making processes. The combination of various methodologies and techniques presented in the paper lays the groundwork for future research and practical applications in predictive analytics and data-driven decision-making in the corporate sector.

REFERENCES

[1] W. F. Abror, M. A. Muslim, D. A. A. Pertiwi, and J. Jumanto, "Combination of weak learner and strong on stacking to increase bankruptcy risk prediction," in *International Conference on Optimization Computer Application and Community Service*, vol. 1, no. 1, 2022, pp. 23–28.

[2] W. F. Abror, A. Alamsyah, and M. Aziz, "Bankruptcy prediction using genetic algorithm-support vector machine (ga-svm) feature selection and stacking," *Journal of Information System Exploration and Research*, vol. 1, no. 2, 2023.

[3] T. M. Alam, K. Shaukat, I. A. Hameed, S. Luo, M. U. Sarwar, S. Shabbir, J. Li, and M. Khushi, "An investigation of credit card default prediction in the imbalanced datasets," *IEEE Access*, vol. 8, pp. 201 173–201 198, 2020.

[4] M. Alfonse and A.-B. M. Salem, "Intelligent model for enhancing the bankruptcy prediction with imbalanced data using oversampling and catboost."

[5] H. Aljawazneh, A. Mora, P. García-Sánchez, and P. Castillo-Valdivieso, "Comparing the performance of deep learning methods to predict companies' financial failure," *IEEE Access*, vol. 9, pp. 97 010–97 038, 2021.

[6] S. Aly, M. Alfonse, and A.-B. M. Salem, "Bankruptcy prediction using artificial intelligence techniques: a survey," in *Digital Transformation Technology: Proceedings of ITAF 2020*. Springer, 2022, pp. 335–360.

[7] ——, "Intelligent model for enhancing the bankruptcy prediction with imbalanced data using oversampling and catboost," *International Journal of Intelligent Computing and Information Sciences*, vol. 22, no. 3, pp. 92–108, 2022.

[8] D. Boughaci and A. A. Alkhawaldeh, "Appropriate machine learning techniques for credit scoring and bankruptcy prediction in banking and finance: A comparative study," *Risk and Decision Analysis*, vol. 8, no. 1-2, pp. 15–24, 2020.

[9] D. Boughaci, A. A. Alkhawaldeh, J. J. Jaber, and N. Hamadneh, "Classification with segmentation for credit scoring and bankruptcy prediction," *Empirical Economics*, vol. 61, pp. 1281–1309, 2021.

[10] Y.-S. Chen, C.-K. Lin, C.-M. Lo, S.-F. Chen, and Q.-J. Liao, "Comparable studies of financial bankruptcy prediction using advanced hybrid intelligent classification models to provide early warning in the electronics industry," *Mathematics*, vol. 9, no. 20, p. 2622, 2021.

[11] S. H. Cho and K.-s. Shin, "Feature-weighted counterfactual-based explanation for bankruptcy prediction," *Expert Systems with Applications*, vol. 216, p. 119390, 2023.

[12] M. Elhoseny, N. Metawa, G. Sztano, and I. M. El-Hasnony, "Deep learning-based model for financial distress prediction," *Annals of Operations Research*, pp. 1–23, 2022.

[13] Y.-C. Hu, "A multivariate grey prediction model with grey relational analysis for bankruptcy prediction problems," *Soft Computing*, vol. 24, no. 6, pp. 4259–4268, 2020.

[14] Y.-C. Hu, P. Jiang, H. Jiang, and J.-F. Tsai, "Bankruptcy prediction using multivariate grey prediction models," *Grey Systems: Theory and Application*, vol. 11, no. 1, pp. 46–62, 2021.

[15] W.-C. Lin, Y.-H. Lu, and C.-F. Tsai, "Feature selection in single and ensemble learning-based bankruptcy prediction models," *Expert Systems*, vol. 36, no. 1, p. e12335, 2019.

[16] H. Mateika, J. Jia, L. Lillard, N. Cronbaugh, and W. Shin, "Fallen angel bonds investment and bankruptcy predictions using manual models and automated machine learning," *arXiv preprint arXiv:2212.03454*, 2022.

[17] M. A. Muslim, Y. Dasril, H. Javed, W. F. Abror, D. A. A. Pertiwi, T. Mustaqim *et al.*, "An ensemble stacking algorithm to improve model accuracy in bankruptcy prediction," *Journal of Data Science and Intelligent Systems*, 2023.

[18] D. L. Olson and B. Chae, "Balancing and variable reduction of firm bankruptcy data," *Journal of Supply Chain Management Science*, vol. 3, no. 1-2, pp. 3–15, 2022.

[19] G. Premalatha, R. Priyanka, and K. Chaitya, "Feature selection for predicting bankruptcy: Comparative analysis," in *2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE, 2023, pp. 1–5.

[20] S. A.-D. Safi, P. A. Castillo, and H. Faris, "Cost-sensitive metaheuristic optimization-based neural network with ensemble learning for financial distress prediction," *Applied Sciences*, vol. 12, no. 14, p. 6918, 2022.

[21] C.-F. Tsai, "Two-stage hybrid learning techniques for bankruptcy prediction," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 13, no. 6, pp. 565–572, 2020.

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

[23] K. El Madou, M. El Merouani, S. Marso, and M. El Kharrim, "High efficacy of handling imbalanced data to predict bankruptcy," in *3rd INTERNATIONAL CONFERENCE ON BIG DATA AND MACHINE LEARNING (BML'22) 23th-24th May 2022, Istanbul, Turkey*, p. 73.

[24] S. Al-Deen Safi, P. Á. Castillo Valdivieso, H. Faris *et al.*, "Cost-sensitive metaheuristic optimization-based neural network with ensemble learning for financial distress prediction," 2022.

[25] C.-F. Tsai, "Feature selection in bankruptcy prediction," *Knowledge-Based Systems*, vol. 22, no. 2, pp. 120–127, 2009.

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25]