# BANKRUPTCY PREDICTION

BY: ARYAN DAHIYA

***Abstract:*** *This study uses a combination of machine learning exploratory data analysis to forecast company insolvency. In order to comprehend data properties and extract pertinent elements, a thorough study of financial indicators was carried out. The collection included financial data for a number of businesses with bankruptcy or non-bankrupt status labels. SMOTE (Synthetic Minority Over-sampling Technique) was used to correct the original imbalance in the dataset. To determine which machine learning algorithm was best for forecasting bankruptcy, several methods were examined. In the end, classifiers such as Decision Trees, Support Vector Machines (SVM), and Logistic Regression were used along with metrics including accuracy, precision, recall, and F1-score to assess the performance of the model. The created model provides financial stakeholders with insightful information to help them evaluate business risk and make wise decisions.*

*Key Words: Machine Learning, Exploratory data analysis, Over-sampling, Recall, F1-score.*

## INTRODUCTION:

Financial analysts and investors rely heavily on bankruptcy prediction in order minimize risk. This research investigates the use of machine learning algorithms to forecast a company's bankruptcy status using particular financial ratios and indicators. The objective is to create a trustworthy model that will help identify possible bankruptcy early so that interested parties can take the appropriate precautions.

### Exploratory Data Analysis (EDA):

It diligently examines over the information to learn more about how the features are distributed, find possible correlations with the goal variable (bankruptcy), and find any problems with the quality of the data.

Feature Engineering:

Considering careful consideration, the algorithm chooses a subset of relevant features that may be useful in anticipating bankruptcy. In order to enhance model performance, it could entail transforming existing features or adding new ones depending on domain expertise.

Class Imbalance Mitigation:

Class imbalance is an issue frequently encountered in real-world datasets, where one class (bankrupt corporations, for example) may be greatly underrepresented in comparison to the other (solvent companies, for example). To make sure the model learns from both classes efficiently, this code takes care of the problem.

Machine Learning Model Training:

Leveraging the prepared data, the code trains numerous machine learning algorithms to create models for bankruptcy prediction. These models are able to forecast new, unobserved companies based on trends seen in the data.

Performance Evaluation:

A thorough evaluation approach was carried out to determine the trained models' efficacy. The algorithm used a number of measures to assess how well the model could correctly identify organizations as either bankrupt or not.

## APPROACH AND OUTCOMES:

The study takes a methodical approach that includes feature engineering, training models, data investigation, and evaluation. Exploratory data analysis was first carried out to comprehend the properties of the data, find possible correlations between variables, and find any problems with the quality of the data. Predictive power was increased by carefully selecting and engineering pertinent financial features. Oversampling approaches were used to overcome the widespread problem of class imbalance, where bankrupt enterprises are typically underrepresented.

Before Jumping into EDA, lets understand what each column means:

- **Borrowing dependency**: Borrowing dependency refers to a company's reliance on borrowed funds, such as loans or credit, to finance its operations or investments. It indicates the extent to which a company utilizes external debt to support its activities rather than relying solely on internal resources.

- **Current Liability to Current Assets**: This ratio compares a company's current liabilities (obligations due within one year) to its current assets (assets expected to be converted into cash within one year). It provides an indication of a company's ability to meet its short-term obligations using its short-term assets. A higher ratio may suggest a greater risk of liquidity issues.

- **Debt ratio %**: The debt ratio is a financial metric that compares a company's total debt to its total assets. It represents the proportion of a company's assets that are financed by debt. A higher debt ratio indicates a higher level of debt relative to assets, which may imply higher financial risk and reduced financial flexibility.

- **Net Income to Stockholder's Equity**: This ratio, also known as return on equity (ROE), measures a company's profitability relative to its shareholders' equity. It indicates how effectively a company generates profit using the shareholders' investment. A higher ratio implies better profitability and efficient use of equity capital.

- **Net Value Per Share (A)**: Net Value Per Share is a measure of a company's net assets (assets minus liabilities) divided by the total number of outstanding shares. It represents the per-share value of a company's net worth or book value.

- **Net profit before tax/Paid-in capital**: This ratio compares a company's net profit before tax to its paid-in capital. It indicates the profitability generated by each unit of capital invested by shareholders.

- **Operating Gross Margin**: Operating gross margin, also known as gross profit margin, measures the profitability of a company's core operations. It is calculated by dividing the gross profit (revenue minus the cost of goods sold) by the revenue. It represents the percentage of revenue that remains after deducting the direct costs associated with producing or delivering goods or services.

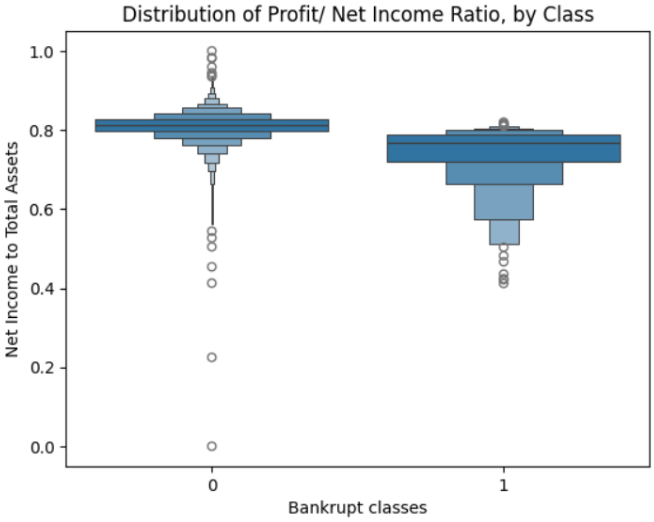
- **Per Share Net profit before tax (Yuan ¥)**: Per Share Net profit before tax is the net profit before tax of a company divided by the total number of outstanding shares. It represents the earnings per share before tax.

- **Persistent EPS in the Last Four Seasons**: Persistent EPS (Earnings Per Share) in the Last Four Seasons refers to the average earnings per share of a company over the past four fiscal quarters. It provides an indication of the company's sustained profitability over a specific period.

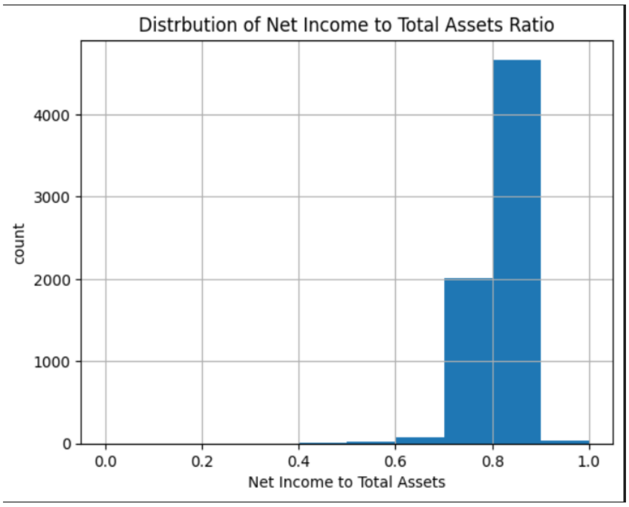
- **ROA(A) before interest and % after tax**: Return on Assets (ROA) measures a company's ability to generate profit from its total assets. ROA(A) before interest and % after tax specifically refers to the return on assets before interest expenses and taxes. It indicates the profitability generated by each dollar of assets, excluding the impact of interest payments and taxes.

- **Working Capital to Total Assets**: This ratio compares a company's working capital (current assets minus current liabilities) to its total assets. It evaluates the proportion of a company's total assets that are funded by its working capital. A higher ratio suggests a higher reliance on short-term assets to finance a company's operations.

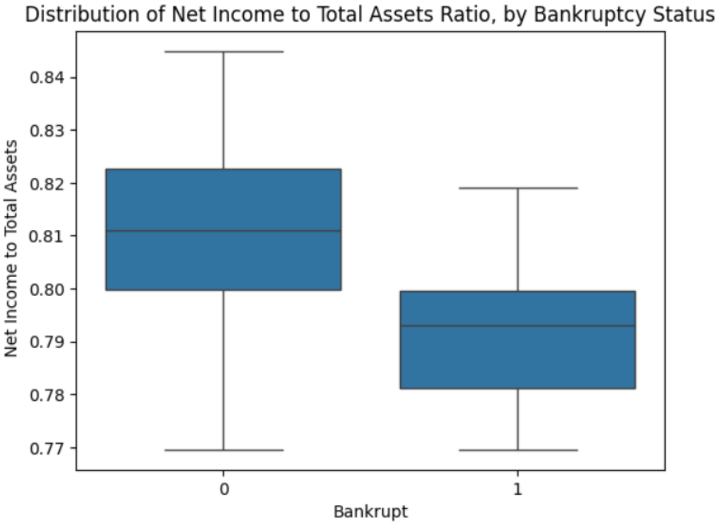
# EXPLORATORY DATA ANALYSIS:



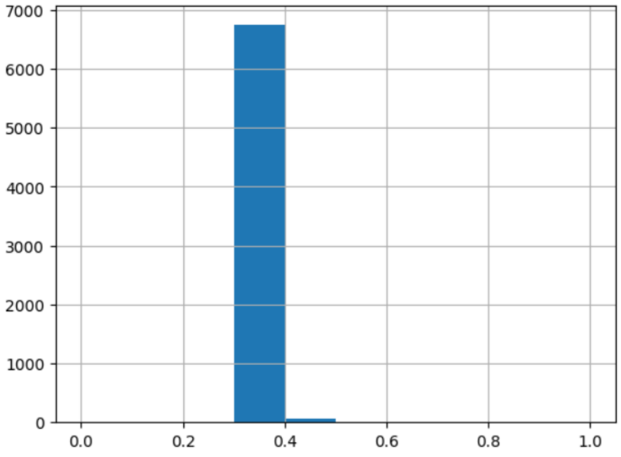
The boxen plot highlights the notable distinctions between the two groups by displaying the distribution of the "Net Income to Total Assets" ratio for non-bankrupt (0) and bankrupt (1) enterprises. Companies that are not bankrupt have a higher median ratio, which is indicative of stronger financial standing, and more variability, as shown by a larger interquartile range (IQR) and a high number of outliers. Conversely, insolvent enterprises exhibit a more compact distribution and a lower median ratio, indicating a lesser level of financial stability and consistency. The clear skewness towards lower values for insolvent businesses highlights even more of their financial difficulties. These findings imply that the "Net Income to Total Assets" ratio, which clearly distinguishes between financially sound and unstable businesses, is a crucial predictor for forecasting bankruptcy.



The distribution of the "Net Income to Total Assets" ratio among companies is depicted by the histogram, which shows a strongly right-skewed trend with most companies concentrated towards the higher end, specifically between 0.7 and 0.8. This suggests that a high ratio is maintained by the majority of businesses, indicating strong financial health. The 0.8 range exhibits the highest frequency, with very few companies falling below 0.6. This suggests a general tendency of strong profitability in relation to total assets. The skewness emphasizes the majority of financially sound businesses, with a small number showing lower ratios that might indicate a higher danger of bankruptcy.

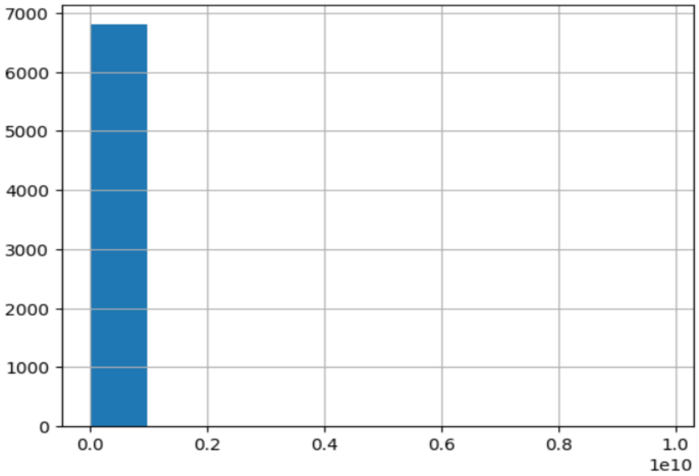


The box plot displays the distribution of the ratio "Net Income to Total Assets" between companies that are (1) bankrupt and those that are not (0). The median ratio of non-bankrupt enterprises is 0.81, and its interquartile range (IQR) is rather narrow, ranging from roughly 0.80 to 0.82. These data points to a consistent pattern of increased profitability relative to total assets. As a result of greater variation in their financial condition, insolvent enterprises have a slightly wider IQR and a lower median ratio of 0.79. The plot highlights the difference in financial stability between non-bankrupt and bankrupt enterprises, with the former often attaining better and more stable profitability ratios than the latter.

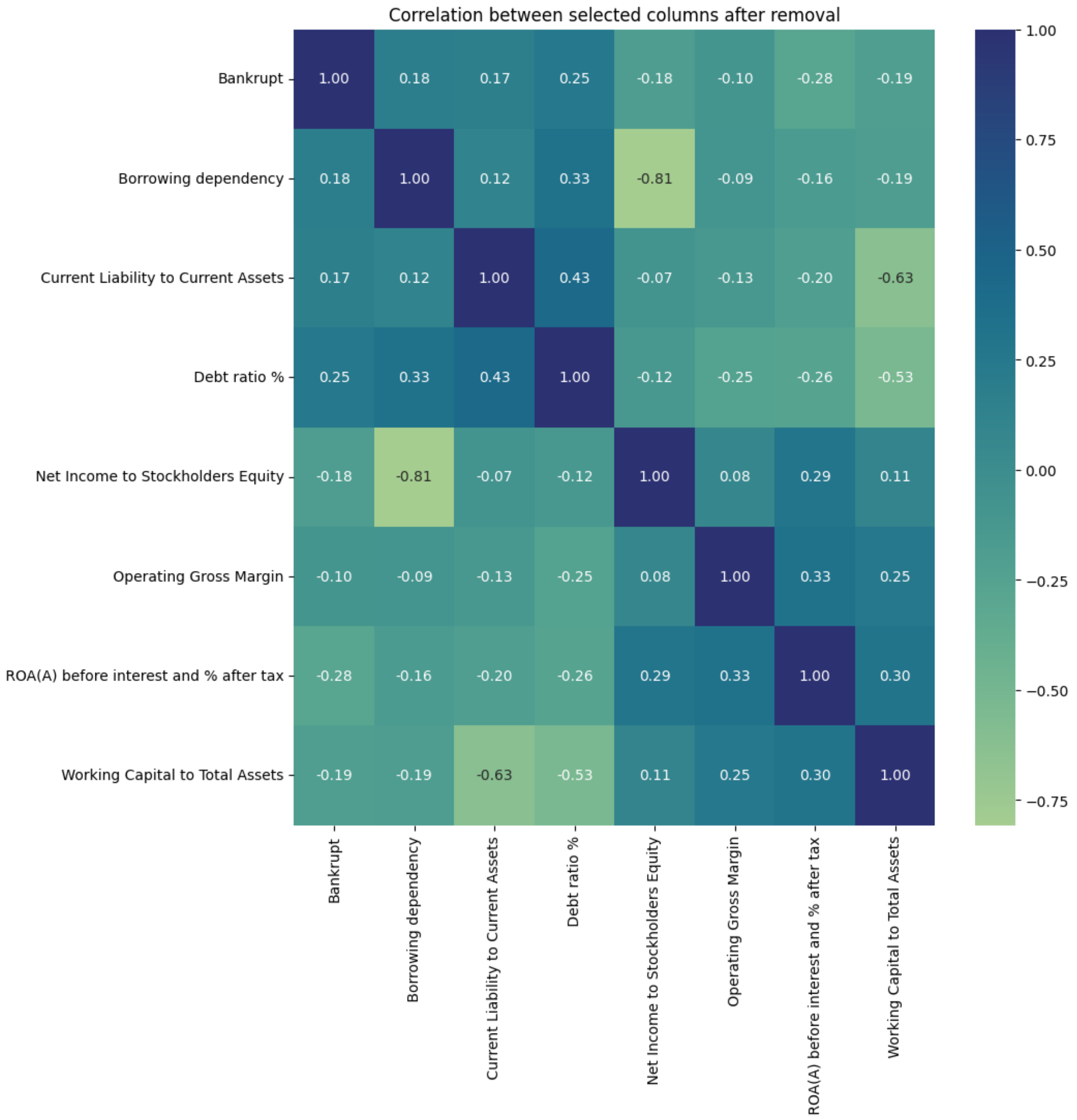


The 'Borrowing dependency' feature in the dataset has a highly skewed distribution, with most of the values clustered around 0.4, according to the histogram. The dataset's skewness indicates that the majority of the companies have a similar borrowing reliance, with only a small number displaying higher or lower values. Machine learning models may have difficulties as a result of this data skewness, which could result in inaccurate or subpar performance.

The dataset's 'Total assets to GNP price' histogram shows a highly skewed distribution, with a long tail stretching towards higher values and the majority of values grouped at zero. This suggests that while a small number of businesses have noticeably larger ratios, the majority of enterprises have relatively modest ratios of total assets to GNP price. A distribution like this might cause bias and impair model performance, which can be problematic for statistical analysis and machine learning.

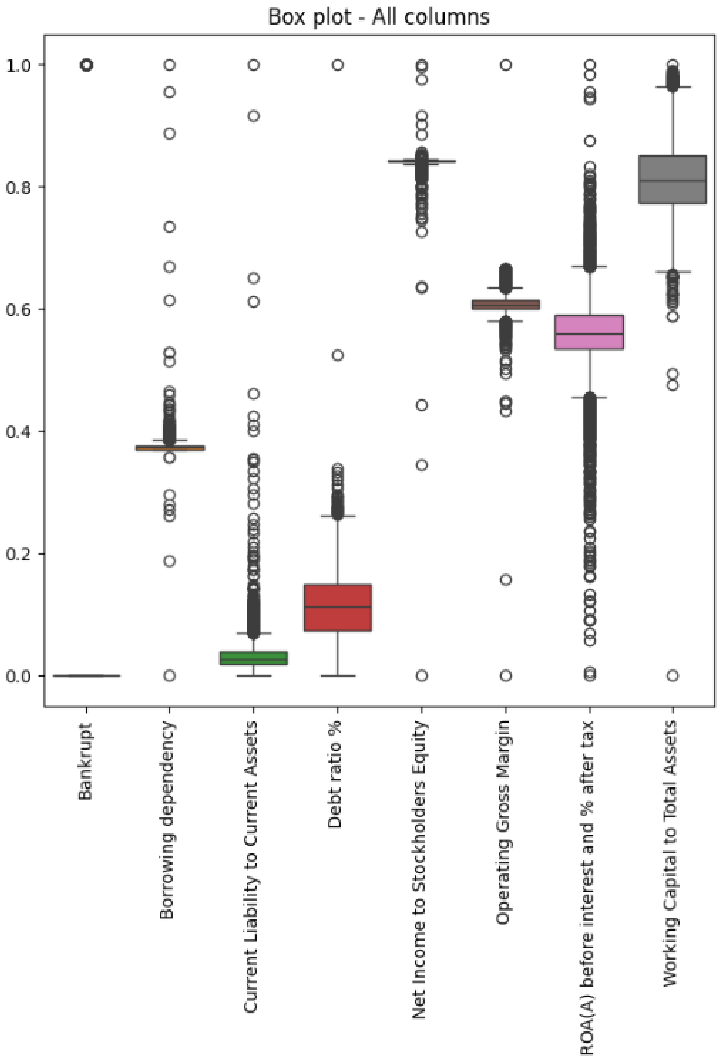


The selected columns for the analysis include 'Bankrupt', 'Borrowing dependency', 'Current Liability to Current Assets', 'Debt ratio %', 'Net Income to Stockholders Equity', 'Net Value Per Share (A)', 'Net profit before tax/Paid-in capital', 'Operating Gross Margin', 'Per Share Net profit before tax (Yuan ¥)', 'Persistent EPS in the Last Four Seasons', 'ROA(A) before interest and % after tax', and 'Working Capital to Total Assets'. All these columns, chosen based on their linear and non-linear relationships, are continuous variables. However, there is a high correlation (|>0.7|) between some columns. Specifically, the columns 'Net Value Per Share (A)', 'Net profit before tax/Paid-in capital', 'Net Income to Stockholder's Equity', 'Persistent EPS in the Last Four Seasons', and 'Per Share Net profit before tax (Yuan ¥)' exhibit high correlations. To mitigate multicollinearity, we will remove these highly correlated columns and then examine the distribution of the remaining data based on the 'Bankruptcy' status.



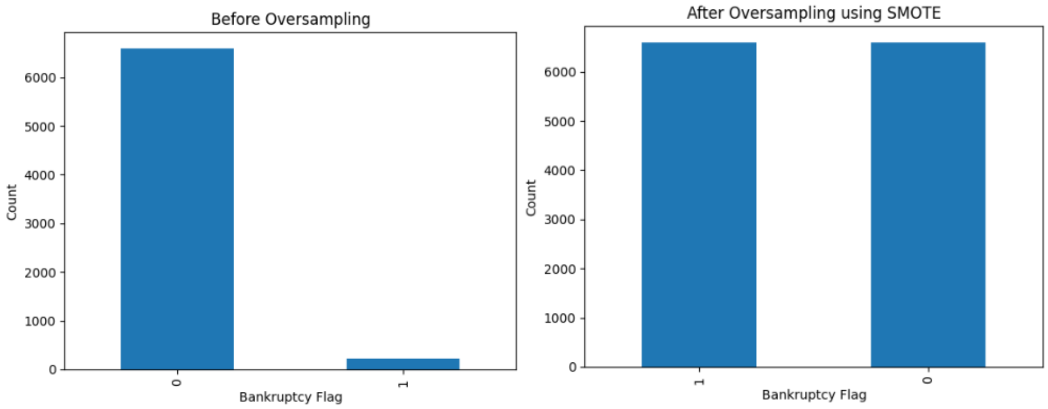
After deleting strongly linked features, the heatmap shows the correlation matrix between the chosen columns. "Bankrupt," "Debt ratio%," "Current Liability to Current Assets," "Borrowing dependency," "Net Income to Stockholders Equity," "Operating Gross Margin," "ROA(A) before interest and % after tax," and "Working Capital to Total Assets" are the remaining columns. The heatmap shows that there may be less chance of multicollinearity because the correlations between these features are now often below the |0.7| threshold. The development of more durable and dependable products will be aided by this enhanced feature set. By concentrating on these factors, we hope to reduce the impact of redundant information and more precisely assess and predict bankruptcy status.

The provided box plot visualizes the distribution of various financial ratios and metrics for the companies in the dataset, grouped by the Bankrupt column. Each box represents the interquartile range (IQR) of the respective feature, with the line inside the box indicating the median value. From the box plot, we can observe that there are significant variations and outliers in some of the financial metrics, such as Debt ratio % and Net Income to Stockholders Equity. These variations could provide insights into the financial health and stability of the companies, and how these metrics differ between bankrupt and non-bankrupt companies. The box plot helps identify the central tendency, spread, and skewness of the financial metrics, which are essential for understanding the financial characteristics of the companies in the dataset.



## MODEL TRAINING:

The code employs a 70-30 train-test split to partition the dataset. This means 70% of the data is allocated for training the machine learning models, while the remaining 30% is held back for evaluation. The train_test_split function from the sklearn.model_selection module is utilized for this purpose, ensuring a random distribution of samples between the training and testing sets.



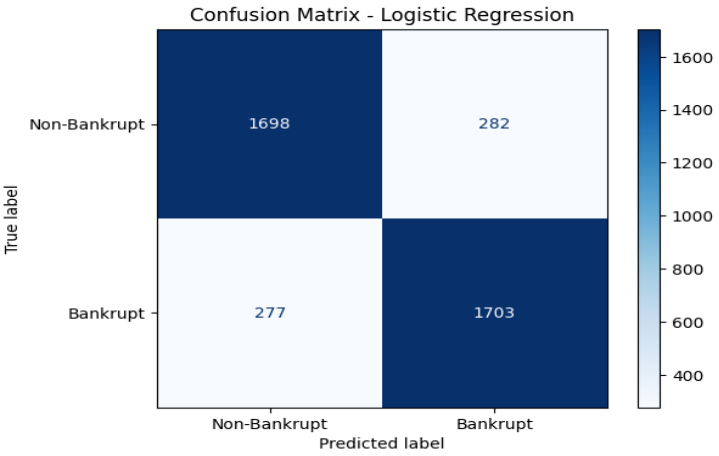
The bar charts effectively visualize the transformation of the dataset from a state of significant class imbalance to a balanced state. Initially, the stark disparity between the number of bankrupt and non-bankrupt instances highlighted the challenge of training a robust model. However, the application of SMOTE oversampling successfully rectified this issue, resulting in a dataset where both classes are represented equally. This balanced representation is crucial for developing a fair and accurate predictive model.

Three primary classification algorithms are implemented and evaluated in the code:

1. **Logistic Regression:** This model is suitable for binary classification problems like bankruptcy prediction. It estimates the probability of a company going bankrupt based on the input features.
2. **Support Vector Machine (SVM):** SVM is known for its effectiveness in classification tasks, particularly when dealing with complex decision boundaries. It aims to find the optimal

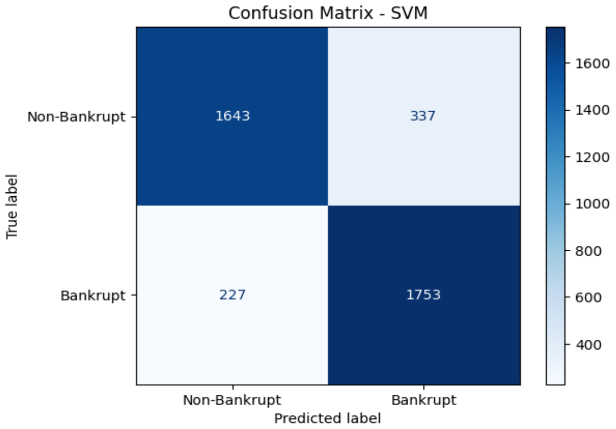
hyperplane that separates the two classes (bankrupt and non-bankrupt).

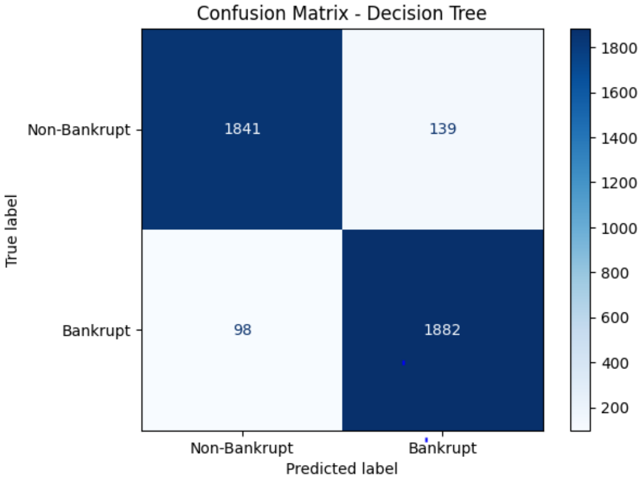
3. **Decision Tree:** This algorithm creates a tree-like model of decisions and their possible consequences, leading to a classification. It can capture non-linear relationships between features and the target variable.



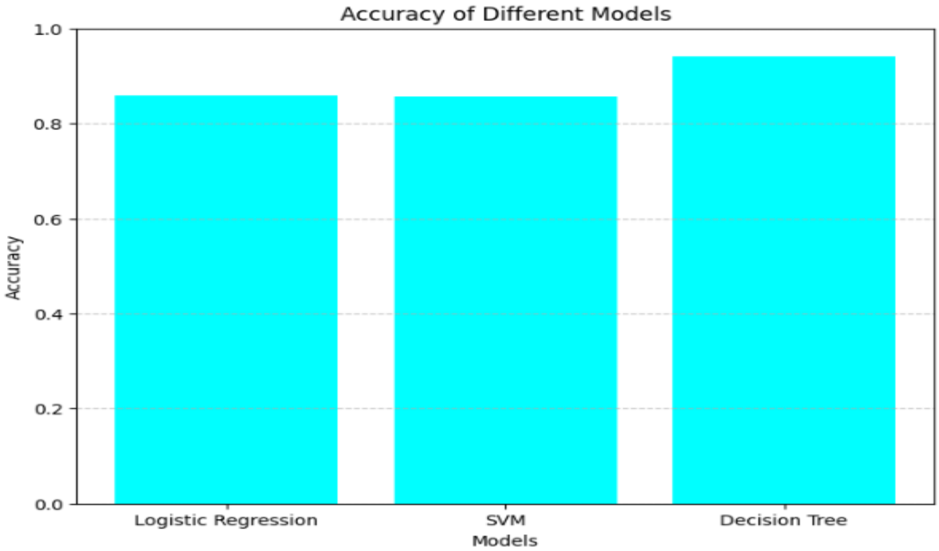
This is the confusion matrix for a bankruptcy prediction logistic regression model. For "Non-Bankrupt" and "Bankrupt" situations, the true labels and predicted labels are represented by two rows and two columns in the matrix. 1698 true non-bankrupt cases were accurately anticipated to be non-bankrupt, while 1703 true bankrupt cases were correctly forecasted to be bankrupt, as indicated by the diagonal elements, which display accurate predictions. Misclassifications are evident in the off-diagonal elements: 277 bankrupt cases were mistakenly forecasted as non-bankrupt, and 282 non-bankrupt cases were mistakenly predicted as bankrupt. With the use of this matrix, we can evaluate the model's effectiveness in identifying organizations as bankrupt or not, highlighting both its accurate and inaccurate predictions.

This is the confusion matrix for a bankruptcy prediction SVM (Support Vector Machine) model. The model's performance in categorizing situations as "Non-Bankrupt" or "Bankrupt" is displayed in the matrix. 1643 non-bankrupt instances were accurately classified, according to the top-left cell; 1753 bankrupt cases, on the other hand, were correctly identified, according to the bottom-right cell. The model incorrectly identified 227 bankrupt cases as non-bankrupt and 337 non-bankrupt cases as bankrupt (bottom-left cell). This image makes it possible to quickly evaluate the model's accuracy, highlighting both its strengths and faults in terms of accurate predictions and misclassifications for each category.





This is the confusion matrix for a bankruptcy prediction Decision Tree model. The model's performance in categorizing situations as "Non-Bankrupt" or "Bankrupt" is displayed in the matrix. 1841 non-bankrupt instances were correctly classified, according to the top-left cell; 1882 bankrupt cases were correctly identified, according to the bottom-right cell. The model incorrectly identified 98 bankrupt cases as non-bankrupt and 139 non-bankrupt cases as bankrupt (bottom-left cell). With relatively few misclassifications in either category, this graphic offers a clear overview of the accuracy of the Decision Tree model and highlights its outstanding performance in properly predicting both bankrupt and non-bankrupt situations.



A bar chart comparing the accuracy of Logistic Regression, SVM, and Decision Tree models is presented. Decision Trees exhibit the highest accuracy, followed by Logistic Regression, with SVM demonstrating the lowest performance.

## CONCLUSION:

This study successfully developed a bankruptcy prediction model using a combination of exploratory data analysis, feature engineering, and machine learning techniques. Addressing class imbalance through oversampling proved crucial for model performance. Among the evaluated models, Decision Trees demonstrated the highest accuracy, indicating their potential for effectively classifying bankrupt and non-bankrupt companies. Ultimately, this model offers a valuable tool for financial analysts and stakeholders to assess company risk and make informed decisions.