# Bankruptcy Prediction Project Report

## Project Overview

The project focused on building and evaluating several machine learning classification models to predict corporate bankruptcy based on various financial indicators. It followed a standard machine learning workflow, encompassing data loading, preprocessing, model training, and evaluation. A significant part of the analysis involved assessing the impact of Principal Component Analysis (PCA) on model performance.

## Data Preprocessing

The initial `data.csv` dataset underwent essential preprocessing steps:

- **Missing Values**: Rows containing any missing values were systematically removed from the dataset.
- **Constant Columns**: Columns exhibiting only a single unique value (i.e., invariant entries) were identified and subsequently dropped, as they lack predictive utility.
- **Feature Scaling**: Features were standardized using `StandardScaler`. This crucial step ensures that all features contribute equitably to the model, preventing features with larger numerical ranges from disproportionately influencing the learning process.

## Model Performance (Without PCA)

The following classification models were evaluated on the preprocessed data, prior to the application of PCA:

- **Logistic Regression**: Achieved high accuracy but showed limitations in precisely identifying bankruptcy cases (low precision and recall for the positive class).
- **Random Forest**: Similar to Logistic Regression, it demonstrated high overall accuracy but struggled with recall for the minority class, though with slightly better precision.
- **Support Vector Machine (SVM)**: Exhibited perfect precision for the positive class but suffered from extremely low recall, indicating a very conservative approach to predicting bankruptcy.

- **K-Nearest Neighbors (KNN)**: Its performance was comparable to Logistic Regression, maintaining high accuracy but with restricted ability to detect true bankruptcy instances.

## Model Performance (With PCA)

The project also investigated the effect of PCA for dimensionality reduction. PCA aims to minimize the number of features while preserving the majority of the data's variance. The models were re-evaluated using the PCA-transformed features:

- **Logistic Regression (with PCA)**: Showed a marginal decrease in accuracy and precision.
- **Random Forest (with PCA)**: Maintained high accuracy, but experienced a notable drop in recall, suggesting potential loss of critical information for bankruptcy identification.
- **Support Vector Machine (with PCA)**: Demonstrated a slight improvement in accuracy, but recall remained very low, indicating that PCA did not substantially enhance the detection of actual bankruptcy cases for SVM.
- **K-Nearest Neighbors (with PCA)**: Its performance closely mirrored that of the non-PCA version.

The results after PCA indicated that while dimensionality reduction can simplify the model, it can also lead to a trade-off in predictive performance, especially concerning recall for the minority class.

## Most Responsible Features for Bankruptcy

**From Random Forest Feature Importances (Top 10):**

- Net Income to Stockholder's Equity
- Persistent EPS in the Last Four Seasons
- Net profit before tax/Paid-in capital
- Equity to Liability
- Net Income to Total Assets
- Net Value Per Share (B)
- Interest Expense Ratio
- Cash/Total Assets
- Degree of Financial Leverage (DFL)
- Borrowing dependency

**From Logistic Regression Coefficients (Absolute, Top 10):**

- Persistent EPS in the Last Four Seasons
- Net Income to Total Assets
- Net Value Per Share (B)
- Operating profit/Paid-in capital
- Operating Profit Rate
- Debt ratio %
- Net worth/Assets
- Cash Flow to Liability
- Total income/Total expense
- Accounts Receivable Turnover

These features consistently highlight that **profitability, debt levels, and efficient asset management** are key financial indicators for predicting bankruptcy.

## Overall Conclusion

The project successfully developed and assessed several machine learning models for bankruptcy prediction, consistently achieving high overall accuracy. However, challenges persist in effectively identifying the minority class (bankruptcy cases), particularly for models like SVM. While PCA offers benefits in dimensionality reduction, its application requires careful consideration to avoid sacrificing critical predictive information. Future endeavors could focus on advanced techniques to address class imbalance and further refine the models' ability to detect bankruptcy with higher recall.