JAI: Jurnal Akuntansi Integratif
Vol. 11 No. 2 (2025): Volume 11 Nomor 2 Oktober 2025, 304-314
ISSN: 2715-0658 [Online]; 2502-5376 [Printed]
10.29080/jai.v11i2.2269

Comparative Analysis of Machine Learning Models in Predicting Corporate Bankruptcy

Alphasyah Lazuardy Sidarta
Accounting Department, School of Accounting, BINUS University, Indonesia
E-mail: alphasyah.sidarta@binus.edu

Abstract

This study aims to compare the performance of several machine learning algorithms Logistic Regression, Decision Tree, Random Forest, and XGBoost in predicting corporate bankruptcy based on financial ratio data, and to provide recommendations regarding their practical applicability. The research employed a dummy dataset sourced from Kaggle, which was preprocessed through cleaning, transformation, and encoding. Data were divided into training (80%) and testing (20%) subsets using stratified sampling to address class imbalance. Model performance was evaluated using accuracy, precision, recall, and F1-score. The results reveal that all models demonstrated high predictive capability, with Random Forest achieving the highest accuracy (96.7%), closely followed by XGBoost (≈96%). Logistic Regression (96%) and Decision Tree (95%) also showed strong results. Profitability and leverage indicators, particularly Net Income to Total Assets and debt ratios, emerged as the most influential predictors. The findings underscore that ensemble tree-based methods offer marginally superior performance due to their ability to capture non-linear interactions, while logistic regression remains valuable for interpretability. The use of a dummy dataset limits the direct generalizability of the findings to real-world financial systems. Moreover, reliance on a single train–test split may overestimate model stability. Future research should employ real-world datasets, apply cross-validation techniques, and explore explainability methods such as SHAP or LIME to enhance transparency. This study contributes by providing a comparative evaluation of machine learning models in bankruptcy prediction, highlighting both accuracy and interpretability trade-offs. The results offer practical insights for auditors, investors, and regulators in selecting predictive tools that balance performance, transparency, and decision-making needs.

Keywords*: Bankruptcy Prediction; Financial Ratios; Machine Learning; Springate Model; Altman Z-Score.*

Introduction

Corporate bankruptcy is a profound and complex economic phenomenon that transcends the simple financial failure of a single enterprise. Instead, it represents a breakdown of organizational resilience, governance integrity, and market trust within broader economic ecosystems (Korol, 2020). The insolvency event is typically catalyzed by a synergistic interplay of endogenous vulnerabilities—such as flawed strategic planning, aggressive risk mismanagement, and ineffective internal controls—and exogenous shocks, including macroeconomic volatility, sector-specific downturns, or unexpected regulatory shifts (Mamatzakis et al., 2015). At the micro-level, bankruptcy manifests as a firm's inability to sustain operations due to persistent losses, severely restricted access to vital capital (Rico & Puig, 2021), or a critical deterioration of reputation among creditors and investors. From a systemic perspective, widespread corporate defaults can trigger domino effects, destabilize labor markets, and fundamentally weaken the stability of financial institutions.

Consequently, the accurate and timely anticipation of corporate failure is not merely a micro-level concern for managers and investors but an imperative for maintaining systemic financial stability and economic health (Toha & Retnaningsih, 2020).

Despite the high urgency of bankruptcy prediction, a central methodological challenge has persisted due to the limitations inherent in traditionally applied approaches. Conventional models, such as the seminal Altman Z-Score and subsequent statistical extensions like discriminant analysis and linear regression, rely fundamentally on linear combinations of selected financial ratios to serve as early warning indicators. While these models retain value for their simplicity and clear interpretability, their predictive efficacy often degrades significantly in complex, heterogeneous, and rapidly evolving business environments. They assume stable, linear relationships between inputs and outcomes—an assumption that rarely holds in contemporary financial markets characterized by deep uncertainty and non-linear interactions among financial, operational, and market variables. This restriction necessitates the adoption of more sophisticated analytical instruments capable of moving beyond linear thresholds and effectively capturing the latent structures, dynamic shifts, and subtle anomalies prevalent in corporate financial data. Machine Learning (ML) has emerged as the definitive methodological response to this requirement, fundamentally transforming the capabilities of risk assessment and decision support across the finance sector (Roscher et al., 2020; Janiesch et al., 2021).

Within the ML paradigm, research efforts in bankruptcy prediction have converged on the comparative performance evaluation of various classification algorithms. Techniques such as Logistic Regression (as a strong linear baseline), Decision Trees, Random Forests, and eXtreme Gradient Boosting (XGBoost) have garnered particular attention due to their proven capabilities in handling complex, high-dimensional, and often imbalanced financial datasets (Nti et al., 2022). Each algorithm presents a unique methodological profile and performance trade-off: Logistic Regression offers high transparency but limited non-linear capacity; Decision Trees provide intuitive rule-based insights but suffer from potential instability; while ensemble methods like Random Forest and the state-of-the-art boosting model, XGBoost, prioritize prediction accuracy and robustness through iterative error optimization. Crucially, while a large volume of literature performs comparative analyses of these ML models (e.g., [Cite relevant comparative studies]), most existing studies suffer from specific limitations. These include focusing primarily on highly regulated, large-firm datasets (e.g., U.S. public firms), utilizing outdated data periods, or neglecting the critical trade-off between complex predictive power and essential model interpretability for regulatory compliance. The clear research gap, therefore, lies in the need for a rigorous, contemporary comparative analysis that specifically evaluates the predictive power, efficiency, and robustness of this essential model spectrum (Linear vs. Tree vs. Ensemble).

Within the domain of bankruptcy prediction, research focus is placed upon the optimal performance of various ML algorithms. Among the wide range of ML techniques, algorithms such as Decision Trees, Random Forests, Logistic Regression, and XGBoost have attracted considerable attention (Nti et al., 2022). Each offers unique advantages and trade-offs. Logistic Regression remains a strong baseline model due to its simplicity and interpretability, though its linear framework restricts its ability to detect complex non-linear patterns. Decision Trees provide an intuitive hierarchical representation of decision rules but are prone to overfitting (Wang, 2017). Random Forests address this weakness by aggregating multiple trees, yielding higher stability and robustness. Meanwhile, XGBoost, a state-of-the-art boosting algorithm, focuses on iterative optimization of prediction errors, excelling in both efficiency and accuracy (Moubayed et al., 2018). The diversity of these algorithms underscores the need for a careful balance between predictive accuracy, model interpretability, and computational efficiency when applied to real-world financial prediction tasks.

Research into the balance of these models is paramount, as the application of ML to corporate bankruptcy prediction carries significant implications from both theoretical and practical standpoints. Theoretically, it contributes to the ongoing shift in financial analytics toward data-driven and non-linear modeling, enriching existing literature on financial distress prediction. Practically, it offers firms, investors, and regulators the opportunity to detect early warning signals of financial vulnerability, thereby enabling timely interventions. By uncovering hidden structures in financial ratios and operational data, ML models facilitate preventive measures before firms reach critical stages of distress. In this sense, ML is not only a predictive tool but also a decision-support system that enhances transparency, accountability, and resilience in financial markets. Hence, comparative empirical studies are required to evaluate the relative performance of these key algorithms within the financial data environment.

Building upon these outlined considerations and research necessity, this study seeks to conduct a comparative analysis of several widely used ML algorithms Logistic Regression, Decision Tree, Random Forest, and XGBoost in predicting corporate bankruptcy. By rigorously evaluating and contrasting their performance using accuracy, precision, recall, and F1-score, the study aims to provide empirical evidence regarding the most effective models for bankruptcy prediction. The findings are expected to enrich academic discourse on financial risk modeling while also offering practical recommendations for stakeholders seeking robust, efficient, and interpretable predictive tools in corporate finance.

Research Method

Time and Location of the Study

This research was carried out over a three-month period, from March to May 2025, allowing sufficient time for data preparation, model construction, evaluation, and validation of results. All stages of data processing and model training were performed online, utilizing the Python programming language within the Google Colab environment. Google Colab was chosen due to its cloud-based computational capacity, accessibility, and compatibility with machine learning libraries, which facilitated efficient experimentation without the need for high-end local infrastructure. The use of Colab also ensured that the workflow was reproducible and collaborative, as codes and outputs could be easily shared and documented. Computations were primarily executed on a personal laptop, which served as the interface for data management, code implementation, and visualization. Although limited in hardware specifications compared to high-performance servers, the integration with Colab's cloud processing allowed the research to overcome computational constraints typically encountered in machine learning experiments.

Data Source

The study employed secondary data, specifically a dummy dataset designed to simulate real-world corporate financial information. This dataset contained multiple financial variables, including the probability indicators that determined whether a company was classified as bankrupt or non-bankrupt. The dataset was sourced from Kaggle, an open-source data science community and platform widely recognized for providing diverse and wellstructured datasets for analytical and experimental purposes. Kaggle was selected not only because of its accessibility but also because it is a trusted repository used by researchers and practitioners for benchmarking machine learning models. Although the dataset does not originate directly from corporate financial statements, it serves as a representative proxy for testing machine learning algorithms under controlled conditions. The dataset's pre-structured format allowed for a streamlined preprocessing stage, including cleaning, transformation, and encoding, which was essential to ensure compatibility with machine learning workflows. Nevertheless, it is important to acknowledge that the reliance on dummy data may limit the direct generalizability of the findings to real-world financial systems; however, it provides a suitable foundation for methodological comparison and exploratory validation of predictive models.

Tools and Materials

This study used Python version 3.10 and the following libraries:

a) Pandas and NumPy for data manipulation.

b) Scikit-Learn for implementing several machine learning algorithms such as Logistic Regression, Decision Tree, and Random Forest. Each of these machine learning algorithms is described below:

1) Logistic Regression

Logistic Regression is a statistical model commonly used for classification problems, particularly binary classification. This model has also been applied across various research domains, such as medical research to analyze risk factors for diseases, and in finance for credit risk and bankruptcy risk assessment, which are useful for risk evaluation and institutional decision-making (Li, 2025).

2) Decision Tree

A Decision Tree is a graph that represents choices and their outcomes in a tree structure. In such a graph there are nodes that represent events or choices, and edges that represent decision conditions. Each tree consists of nodes and branches. Each node contains a representation of the attributes of the group to be classified, while each branch represents the values that the node can take (Mahesh, 2020).

3) Random Forest

Random Forest (RF) is a popular machine learning technique in the field of data mining. This technique operates under supervised learning and is used to make predictions about future trends by analyzing features of predictive factors that are useful in future decision-making. The model is constructed by analyzing the characteristics of the variables used for prediction, producing hypotheses that can be empirically tested. Random Forest is built by generating multiple Decision Trees through random sampling of the data using bootstrap sampling and by selecting inputs at random. One advantage of Random Forest is its high accuracy compared to other approaches such as bagging and boosting. This technique also works effectively on large databases and can handle many variables, allowing the analysis of thousands of input variables without needing to remove them (Salman et al., 2024).

XGBoost for boosting models

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable implementation of the boosting technique, which operates by sequentially combining multiple weak learners, typically decision trees, into a single strong predictive model. The core principle of boosting lies in the idea that when the classification performance of one weak learner is inadequate, a new learner is trained to correct the errors or residuals of the previous models. Over successive iterations, the model gradually improves its predictive power by minimizing bias and optimizing the overall performance. Unlike traditional boosting algorithms, XGBoost introduces several technical improvements such as regularization parameters (L1 and L2) to prevent overfitting, shrinkage to reduce the influence of each tree, and advanced handling of missing values, all of which contribute to its superior generalization ability.

In practical applications, XGBoost has become one of the most widely adopted algorithms in both academic research and industry due to its computational efficiency and robustness in handling large, high-dimensional datasets. For financial risk assessment and bankruptcy prediction, these advantages are particularly relevant because financial data are often noisy, imbalanced, and influenced by complex, non-linear interactions. The algorithm's ability to capture subtle patterns within such data enables practitioners to obtain a more reliable and accurate classification of corporate bankruptcy risk compared to conventional models. Furthermore, the flexibility of XGBoost allows integration with various evaluation metrics (such as precision, recall, and F1-score), making it highly adaptable to decision-making contexts where the cost of misclassification is substantial. Consequently, applying XGBoost to bankruptcy prediction not only produces superior composite scores relative to other models but also provides decision-makers such as auditors, investors, and regulators with a practical and powerful tool for anticipating financial distress in dynamic business environments.

Research Procedure

The research procedures consist of the following stages:

Data Collection and Cleaning

The dataset obtained for this study was initially provided in a raw format and then converted into CSV to ensure compatibility with Python-based data analysis tools. Data cleaning was considered a critical stage, as the quality of the input data directly affects the robustness of the machine learning models. During this process, missing values were identified and removed to avoid biased estimations, duplicate entries were eliminated to prevent redundancy and overrepresentation of certain patterns, and categorical variables were encoded into numerical formats to enable their proper processing by the algorithms. These preprocessing steps are essential not only to improve data integrity but also to ensure consistency across the training and testing phases, ultimately enhancing the reliability of the predictive results.

Dataset Splitting

After the cleaning process, the dataset was divided into training and testing subsets, with 80% allocated for training and 20% reserved for testing. This ratio was chosen to provide the models with sufficient data to learn underlying patterns while retaining enough samples to objectively assess predictive performance. Stratified sampling was applied to maintain the same proportion of bankrupt and non-bankrupt firms across both subsets, thereby addressing potential class imbalance and ensuring that the evaluation metrics would not be skewed toward the majority class. Such a technique is particularly important in bankruptcy prediction tasks, where the distribution between positive and negative classes is often highly uneven.

Implementation of Machine Learning Models

This study implemented four machine learning algorithms, each selected to capture different aspects of the predictive problem. Logistic Regression was used as the baseline model due to its interpretability and established application in financial risk prediction. The Decision Tree model was included to provide an easily visualized, rule-based structure that mimics human reasoning, though it is prone to overfitting when used in isolation. Random Forest, as an ensemble extension of Decision Trees, was employed to overcome overfitting by aggregating the results of multiple randomly generated trees, thus yielding higher accuracy and stability. Finally, XGBoost was implemented as a state-of-the-art gradient boosting method that incrementally optimizes prediction errors, offering superior efficiency and performance in large-scale structured datasets. The combination of these models allowed for a comprehensive comparison between linear, rule-based, ensemble, and boosting approaches in bankruptcy prediction.

Model Performance Evaluation

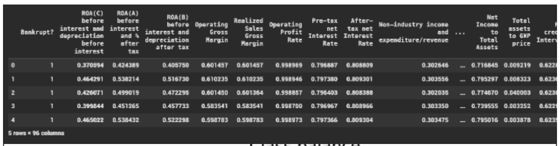
To evaluate the predictive capability of the models, four widely used performance metrics were applied: accuracy, precision, recall, and F1-score. Accuracy provided an overall measure of correctly classified cases, while precision measured the proportion of firms predicted as bankrupt that were actually bankrupt an essential factor for minimizing false alarms in financial contexts. Recall, on the other hand, captured the ability of the models to identify bankrupt firms correctly, which is crucial for early warning systems that aim to prevent financial losses. The F1-score, as a harmonic mean of precision and recall, offered a balanced metric in cases where trade-offs between false positives and false negatives must be considered. Together, these metrics provided a multi-dimensional perspective on model performance, ensuring that the evaluation was not solely focused on overall accuracy but also accounted for the practical implications of misclassification in bankruptcy prediction.

Results

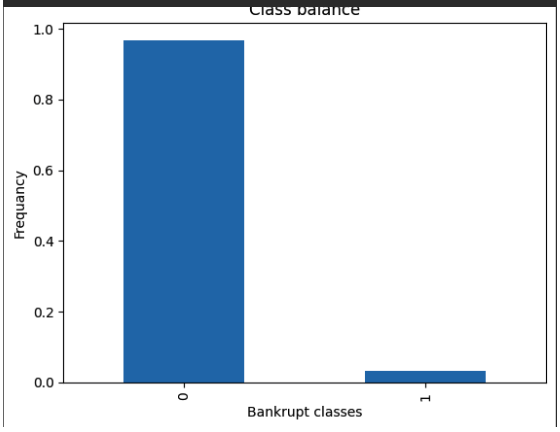
Exploratory Data Analysis and Feature Overview

Figure 1. Class Balance

All preprocessing and initial exploratory analyses were conducted after importing the required Python libraries. The dataset was loaded from the CSV file and examined for class balance, missing values, and basic distributional properties and in Figure 1 displays the relative frequencies of the two bankruptcy classes and motivated the use of stratified sampling during the train–test split to preserve class proportions during model evaluation.



```
                  ROA(C)      ROA(A)        ROA(B)                                                                                                            Net      Total      
                  before      before        before   Operating   Realized  Operating   Pre-tax    After-  Non-industry income         Income     assets     
             interest and   interest  interest and       Gross      Sales     Profit       net   tax net                 and  ...       to   to GNP   cred
             depreciation      and %  depreciation      Margin      Gross       Rate  Interest  Interest                                 Total    price  Inter
   Bankrupt?       before      after     after tax                 Margin                 Rate      Rate  expenditure/revenue           Assets            
                 interest        tax                                                                                                                      
0          1     0.370594   0.424389      0.405750    0.601457   0.601457   0.998969  0.796887  0.808809             0.302646  ...  0.716845  0.009219  0.622
1          1     0.464291   0.538214      0.516730    0.610235   0.610235   0.998946  0.797380  0.809301             0.303556  ...  0.795297  0.008323  0.623
2          1     0.426071   0.499019      0.472296    0.601450   0.601364   0.998857  0.796403  0.808388             0.302035  ...  0.774670  0.040003  0.623
3          1     0.399844   0.451265      0.457733    0.583541   0.583541   0.998700  0.796967  0.808966             0.303350  ...  0.739555  0.003252  0.622
4          1     0.465022   0.538432      0.522298    0.598783   0.598783   0.998973  0.797366  0.809304             0.303475  ...  0.795016  0.003878  0.623
5 rows × 96 columns
```



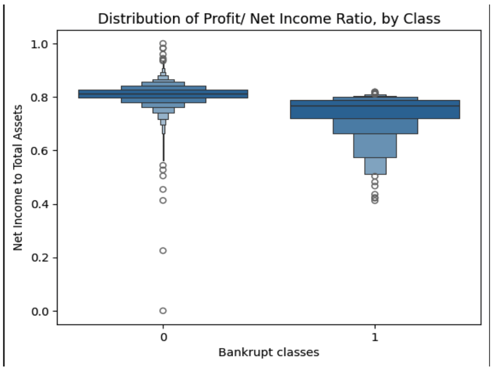


Figure 2. Distribution of Net Income Ratio

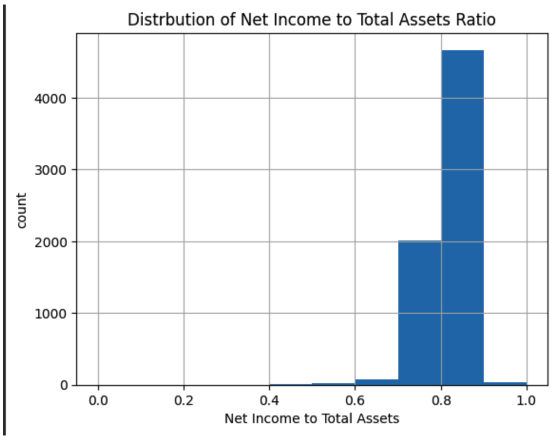
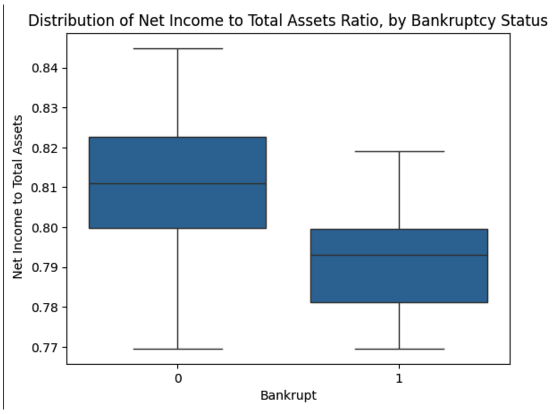


Figure 3. Distribution of Net Income to Total Assets Ratio

Figure 2 presents a boxplot of the ratio Net Income to Total Assets stratified by bankruptcy status. The graphical summary shows differences in central tendency and dispersion between the two groups: companies labelled as bankrupt tend to have lower median values and greater dispersion in this ratio compared with non-bankrupt firms. Complementary to this, the histogram in Figure 3 (and the grouped histogram in Figure 4) illustrates the overall distribution and skewness of the Net Income to Total Assets ratio; bankrupt companies display a distribution shifted toward lower (and negative) values, whereas non-bankrupt firms concentrate around higher positive ratios. Together, these visualizations indicate that profitability-related ratios are potentially discriminative for bankruptcy classification.

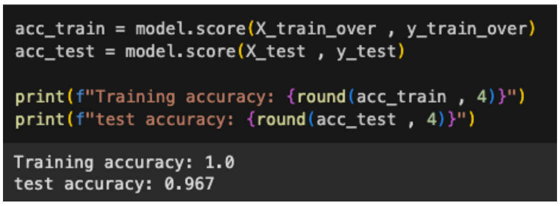
Figure 4. Net Income to Total Assets by Bankuptcy Status



Model Performance

Four machine learning algorithms were trained and evaluated using an 80:20 stratified train–test split. Performance was assessed primarily by accuracy, and the models produced the following test accuracies:

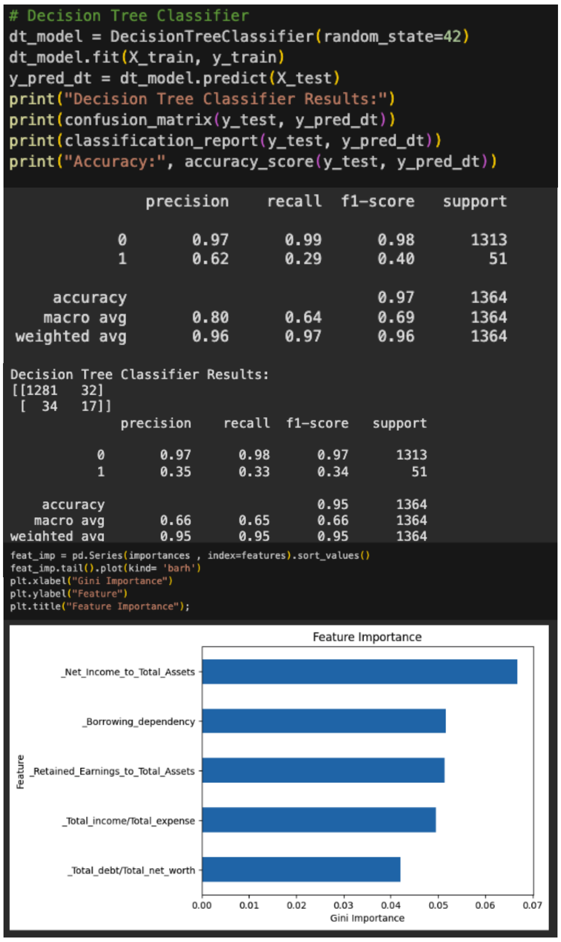
- Random Forest: 0.967 (96.7%). (Detailed RF evaluation in Figure 5).
- Decision Tree: 0.95 (95%). (Figure 6).
- Logistic Regression: 0.96 (96%). (Figure 7).
- XGBoost: 0.96 (96%). (Figure 7).



```
acc_train = model.score(X_train_over , y_train_over)
acc_test = model.score(X_test , y_test)

print(f"Training accuracy: {round(acc_train , 4)}")
print(f"test accuracy: {round(acc_test , 4)}")
```

```
Training accuracy: 1.0
test accuracy: 0.967
```

Figure 5. Random Forest



```
# Decision Tree Classifier
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(X_train, y_train)
y_pred_dt = dt_model.predict(X_test)
print("Decision Tree Classifier Results:")
print(confusion_matrix(y_test, y_pred_dt))
print(classification_report(y_test, y_pred_dt))
print("Accuracy:", accuracy_score(y_test, y_pred_dt))
```

```
              precision    recall  f1-score   support

           0       0.97      0.99      0.98      1313
           1       0.62      0.29      0.40        51

    accuracy                           0.97      1364
   macro avg       0.80      0.64      0.69      1364
weighted avg       0.96      0.97      0.96      1364
```

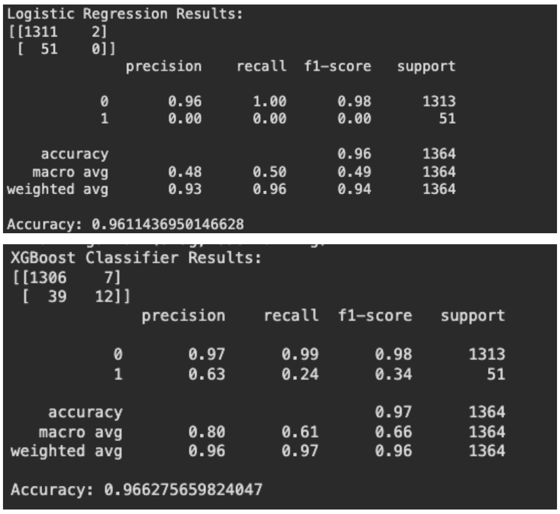
```
Decision Tree Classifier Results:
[[1281   32]
 [  34   17]]
              precision    recall  f1-score   support

           0       0.97      0.98      0.97      1313
           1       0.35      0.33      0.34        51

    accuracy                           0.95      1364
   macro avg       0.66      0.65      0.66      1364
weighted avg       0.95      0.95      0.95      1364
```

```
feat_imp = pd.Series(importances , index=features).sort_values()
feat_imp.tail().plot(kind= 'barh')
plt.xlabel("Gini Importance")
plt.ylabel("Feature")
plt.title("Feature Importance");
```

Feature Importance
_Net_Income_to_Total_Assets
_Borrowing_dependency
_Retained_Earnings_to_Total_Assets
_Total_income/Total_expense
_Total_debt/Total_net_worth
Feature
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Gini Importance

Figure 6. Decision Tree

In addition to accuracy, the study evaluated models using precision, recall, and F1-score (used during the model-performance evaluation stage described in the Methods). The ensemble tree-based methods (Random Forest and XGBoost) achieved the highest overall accuracy on the holdout set, with Random Forest narrowly outperforming XGBoost by a small margin.



```
Logistic Regression Results:
[[1311    2]
 [  51    0]]
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      1313
           1       0.00      0.00      0.00        51

    accuracy                           0.96      1364
   macro avg       0.48      0.50      0.49      1364
weighted avg       0.93      0.96      0.94      1364

Accuracy: 0.9611436950146628
```

```
XGBoost Classifier Results:
[[1306    7]
 [  39   12]]
              precision    recall  f1-score   support

           0       0.97      0.99      0.98      1313
           1       0.63      0.24      0.34        51

    accuracy                           0.97      1364
   macro avg       0.80      0.61      0.66      1364
weighted avg       0.96      0.97      0.96      1364

Accuracy: 0.966275659824047
```

Figure 7. Logistic Reggresion & XGBoost

Discussion

Interpretation of The Exploratory Analysis

The exploratory visualizations provide strong evidence of systematic financial differences between bankrupt and non-bankrupt companies, particularly with respect to profitability and leverage indicators. The lower median and greater dispersion of the Net Income to Total Assets ratio among bankrupt firms (Figures 3–5) are consistent with established theories in financial distress and bankruptcy prediction literature, which emphasize declining profitability and increased volatility in earnings as early warning signals of failure. Firms experiencing persistent losses or unstable earnings streams are generally less able to cover their fixed obligations, leading to heightened default risk. These findings also highlight the presence of heterogeneity within bankrupt firms some exhibit extremely negative profitability, while others fluctuate widely suggesting that financial distress does not follow a uniform trajectory but rather emerges from a combination of poor performance, inconsistent cash flows, and excessive leverage. Taken together, these patterns support the hypothesis that profitability and indebtedness ratios are not only descriptive markers of firm health but also carry substantial predictive power for classification models.

Model Comparison and Implications

The comparative analysis of models demonstrates that tree-based ensemble methods offered the most robust predictive performance. Random Forest recorded the highest accuracy (96.7%), closely followed by XGBoost (≈96%), which validates prior empirical findings that ensemble learning can effectively balance bias and variance in classification tasks. The superiority of these methods over a single Decision Tree (95%) reflects two important dynamics: first, ensembles reduce the risk of overfitting by aggregating predictions from multiple trees, thereby stabilizing outcomes; second, they capture non-linear relationships and variable interactions more effectively than individual models. Interestingly, Logistic Regression also attained a strong accuracy (96%), suggesting that linear approximations of financial ratios already explain a considerable proportion of bankruptcy risk. However, the marginal gains achieved by ensemble models point to the existence of complex, non-linear effects that linear models cannot fully account for, reinforcing the value of advanced machine learning techniques in this domain.

From a practical perspective, these findings carry significant implications for various stakeholders. For financial institutions and regulators, the high accuracy of Random Forest and XGBoost implies that machine learning can be deployed as an effective early-warning system to identify at-risk firms before financial distress becomes irreversible. These tools could complement traditional credit risk assessments by offering more dynamic and data-driven evaluations. On the other hand, Logistic Regression remains appealing in scenarios where interpretability and transparency are paramount, such as regulatory audits or internal risk reporting, since its coefficient estimates can be directly linked to economic meaning. Consequently, the choice of model may depend less on absolute accuracy which is consistently high across methods and more on contextual priorities such as interpretability, scalability, and resource availability.

Why The Identified Features Matter

The most important features identified by the models have clear economic rationale:

- Net Income to Total Assets measures operating profitability relative to asset base; persistently low or negative values are an immediate signal of solvency risk.
- Borrowing dependency and Total Debt / Total Net Worth capture leverage and the firm's dependence on external financing higher leverage increases bankruptcy vulnerability.
- Retained Earnings to Total Assets reflects the cumulative profitability retained in the firm and thus its capacity to absorb losses.
- Total Income / Total Expense is a short-term profitability margin that signals whether core operations can cover ongoing costs.

These variables jointly provide a balanced view of short-term operational viability and longer-term capital structure health, which explains their predictive strength.

Conclusion

This study evaluated the predictive capacity of several machine learning algorithms Random Forest, XGBoost, Logistic Regression, and Decision Tree for corporate bankruptcy classification using financial-ratio features. Exploratory analysis confirmed that profitability and leverage measures, specifically Net Income to Total Assets, borrowing-dependency, and leverage ratios, are the most informative predictors, aligning with established financial theory on solvency.

All tested models demonstrated strong performance, with Random Forest achieving the highest test accuracy (96.7%), closely followed by XGBoost (≈96%) and Logistic Regression (96%). The results suggest that while linear models effectively capture the main bankruptcy risk signal, ensemble tree-based methods provide marginal, but valuable, improvements by leveraging non-linear interactions within the financial data.

In practical terms, the findings indicate that a relatively small set of financial ratios can effectively predict bankruptcy risk, and the optimal algorithm choice should be aligned with organizational priorities, whether emphasizing accuracy (e.g., Random Forest/XGBoost) or interpretability (e.g., Logistic Regression). The models developed herein offer a solid foundation for early warning systems for financial distress. Ultimately, this research demonstrates the significant potential of combining financial ratio analysis with machine learning to enhance predictive accuracy and support stakeholders in risk management.

Limitations

Several limitations warrant mention and should be considered when interpreting the findings of this study. First, the use of a dummy dataset obtained from an online repository (Kaggle), while useful for methodological testing, limits the generalizability of the results, as it may not fully reflect the complexities, inconsistencies, and distributional nuances inherent in real-world corporate financial data. Consequently, the reported predictive performance may differ significantly when applied to authentic financial records from diverse sectors. Second, the model evaluation relies on a single train–test split. This choice, while sufficient for preliminary validation, may overestimate model stability; future research should incorporate more rigorous procedures like k-fold cross-validation or repeated resampling to better assess variance and strengthen the robustness of the evaluation. Third, model explainability presents an inherent challenge, particularly for ensemble methods like Random Forest and XGBoost. The reliance on global feature importance scores limits transparency by failing to provide case-specific justifications for individual predictions. Advanced interpretability techniques, such as SHAP and LIME, are recommended for future studies to enhance the trustworthiness of the application in high-stakes financial contexts. Finally, issues of class imbalance and label quality in the source dataset may influence model calibration and real-world applicability. Bankruptcy datasets are typically imbalanced, leading to potential model bias, and the accuracy of labels may be questionable due to varying definitions of bankruptcy. Future research should prioritize high-quality, domain-specific datasets with balanced representations, ideally verified by authoritative sources.

References

Adi, E., Anwar, A., Baig, Z., & Zeadally, S. (2020, May 11). Machine learning and data analytics for the IoT. *Springer Nature*, *32*, 16205-16233. https://doi.org/10.1007/s00521-020-04874-y

Guan, Y., & Zong, Z. (2024). Corporate Financial Risk Identification and Operation Control Analysis for XGBoost Modeling. *Applied Mathematics and Nonlinear Science*, *vol. 9*, 1-16. https://www.researchgate.net/publication/382553924_Corporate_Financial_Risk_Identification_and_Operation_Control_Analysis_for_XGBoost_Modeling

Janiesch, C., Zschech, P., & Heinrich, K. (2021, April 8). Machine learning and deep learning. *Springer Nature*, *31*, 685-695. https://doi.org/10.1007/s12525-021-00475-2

Korol, T. (2020, April 28). VILNIUS TECH Journals. *Long-term risk class migrations of non-bankrupt and bankrupt enterprises*, *21*(3), 783-804. https://doi.org/10.3846/jbem.2020.12224

Li, X. (2025). improved Logistic Regression Model Based on Resampling Techniques. *Highlights in Science, Engineering and Technology*, *vol. 136*, 28-36. https://www.researchgate.net/publication/390363183_Improved_Logistic_Regression_Model_Based_on_Resampling_Techniques

Mackenzie, A. (2015, June 16). The production of prediction: What does machine learning want? *European Journal of Cultural Studies*, *18*(4-5), 429-445. https://doi.org/10.1177/1367549415577384

Mahesh, B. (2020). Machine Learning Algorithm - A Review. *International Journal of Science and Research*, *vol. 9*, 381-386. 10.21275/ART20203995

Mamatzakis, E., Matousek, R., & Vu, A. N. (2015, April 20). What is the impact of bankrupt and restructured loans on Japanese bank efficiency? *Journal Of Banking & Finance*, *72*, 187-202. https://doi.org/10.1016/j.jbankfin.2015.04.010

Moubayed, A., Injadat, M., Nassif, A. B., Lutfiyya, H., & Shami, A. (2018, July 23). E-Learning: Challenges and Research Opportunities Using Machine Learning & Data Analytics. *IEEE Xplore*, *6*, 39117-39138. 10.1109/ACCESS.2018.2851790

Nti, I. K., Quarcoo, J. A., Aning, J., & Fosu, G. K. (2022, January 25). A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. *IEEE Xplore*, *5*(2), 81-97. 10.26599/BDMA.2021.9020028

Qin, S. J., & Chiang, L. H. (2019, July 12). Advances and opportunities in machine learning for process data analytics. *Computers & Chemical Engineering*, *126*, 465-473. https://doi.org/10.1016/j.compchemeng.2019.04.003

Rico, M., & Puig, F. (2021, February 16). Successful turnarounds in bankrupt firms? Assessing retrenchment in the most severe form of crisis. *BRQ Business Research Quarterly*, *24*(2), 114-128. https://doi.org/10.1177/2340944421994117

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020, February 24). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Xplore*, *8*, 42200 - 42216. 10.1109/ACCESS.2020.2976199

Salman, H. A., Kalakech, A., & Steiti, A. (2024, June 8). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 69-79. https://www.researchgate.net/publication/382419308_Random_Forest_Algorithm_Overview

Tanwar, S. (2024). Machine Learning. *Computational Science and Its Applications*, *1*, 30. 9781003347484

Toha, K., & Retnaningsih, S. (2020, March). Legal Policy Granting Status of Fresh Start to the Individual Bankrupt Debtor in Developing the Bankruptcy Law in Indonesia. *Academic Journal of Interdisciplinary Studies*, *9*(2). https://doi.org/10.36941/ajis-2020-0033

Vasudevan, S. K., Dantu, N. V., Pulari, S. R., & Murugesh, T. S. (2023). What is Machine Learning? *Taylor & Francis Group*, *1*, 16. 9781003393122

Wang, L. (2017, July 24). Data Mining, Machine Learning and Big Data Analytics. *SciEP*, *4*(2), 55-61. 10.12691/iteces-4-2-2

Xu, C., & Zhang, H. (2021, January 14). Real earning management in bankrupt firms. *Journal of Corporate Accounting & Finance*, *32*(2), 22-38. https://doi.org/10.1002/jcaf.22483