Predictive Modeling for Loan Eligibility Assessment: A Comparative Study of Logistic Regression, Random Forest, and Support Vector Machine with Detailed Oversampling

Jamel D. Pandiin, Junrie B. Matias,
Caraga State University - Main Campus, Butuan City, Philippines
jamel.pandiin@carsu.edu.ph

Abstract – This study compares the predictive modeling techniques for loan eligibility assessment, comparing Logistic Regression, Random Forest, and Support Vector Machine (SVM) with detailed oversampling and feature selection methods. Using a Kaggle dataset, various feature selection techniques, including Correlation-Based Selection, Recursive Feature Elimination (RFE), and Lasso Regression was employed for feature selection before applying it to three classifiers: Random Forest, Logistic Regression, and Support Vector Machine (SVM), were optimized through Genetic Algorithms (GA). Performance metrics, including accuracy, precision, recall, and F1-score, alongside cross-validation, were employed for model evaluation. Random Forest achieved the highest performance with an accuracy of 85%, precision of 86%, recall of 84%, and F1-score of 85%. Cross validation results for Random Forest averaged 92%, demonstrating consistent robustness. Feature importance analysis identified Credit History (26.8%), Applicant Income (19.7%), and Loan Amount (19.2%) as the most influential factors, while demographic attributes like Gender and Education had minimal impact. SVM excelled in recall (99%) but showed moderate accuracy (71%) and lower precision (63%), reflecting challenges in minimizing false positives. Logistic Regression exhibited consistent yet lower accuracy (67%) and struggled to model complex relationships in the dataset. The findings highlight Random Forest's strength in delivering balanced predictions, making it the most suitable model for fairness and risk management in loan approvals. Practical deployment via a user-friendly web application demonstrated the usability of machine learning models for operational efficiency in financial institutions. This research advocates for integrating Genetic Algorithms with machine learning for enhanced predictive modeling, ensuring precise, efficient, and fair decision making processes.

Keywords – Predictive Modeling, Loan Eligibility, Genetic Algorithms, Feature Selection.

1 Introduction

Machine learning teaches machines to handle information more efficiently [1]. It is a form of artificial intelligence that enables a computer program to learn from prior tasks. Analyzes data, detects patterns, and requires minimum human participation [2]. Machine learning investigates designing and implementing algorithms capable of making data predictions. It is used to create programs with tuning parameters that are

subsequently adjusted to improve performance by responding to previous data [3]. Over the last few decades, the financial services industry has seen significant developments. On the one hand, the scope of activities in the financial industry has expanded dramatically, encompassing a wide range of new banking, investment, and insurance products, as well as new financing instruments and corporate finance methods [4][5]. Loan lending has played a significant role in the daily lives of both businesses and individuals. With the ever-increasing competitiveness in the financial sector and many financial restraints, accepting loans has become inescapable [6].

Banks provide various products in our banking system, but their primary source of income is credit lines. As a result, they are likely to benefit from the interest in their loans. Loans, or whether clients repay or fail on their loans, impact a bank's profits and losses [7]. People desire to apply for online loans since data grows daily due to digitalization in the banking industry. Artificial intelligence (AI), a common approach to information exploration, has received increased attention. Individuals from diverse businesses use it. AI computations will take care of the issues based on industry knowledge. Banks need help with approving loans. Most banks benefit from loans but selecting suitable consumers from a pool of applicants is hazardous. One error might result in a significant loss for a bank [8].

Several banks and lending companies, tiny ones, still need to adapt their modernization processes. They still use manual methods to choose the approval and borrower data. With the presence of machine learning and AI, we will explore Genetic algorithms and choose among the best classifiers to combine with GA-named Logistic Regression, Random Forest, and SVM to test the accuracy level of their performance, which is fit for predicting the borrower's approval. Evaluate the efficiency models, develop, validate, and optimize machine learning models through rigorous training, fine-tuning, and evaluation using relevant performance metrics. Deploy the model using web applications and implement a user-friendly web application that leverages trained models, enabling seamless data input and accurate user prediction output.

2 Related Literature

Comparative study on binary classification for loan eligibility prediction evaluates five machine learning algorithms SVM, Logistic Regression, KNN, Decision Tree, and Stochastic Gradient Descent highlighting their strengths and limitations [9]; however, it does not explore advanced optimization techniques such as Genetic Algorithms (GA) for feature selection, potentially limiting model performance, and applies SMOTE only to the validation dataset, leaving the training data imbalanced, which could impact real-world applicability [10]. Similarly, Mnkandla et al. (2024) develop a machine learning-based loan eligibility system tailored for the African financial sector, achieving an 80% accuracy with a Logistic Regression model [11] but failing to implement advanced feature selection methods like GA or Lasso regularization and overlooking class imbalance, which can bias predictions [12]. Likewise, propose an AI-driven machine learning application for credit eligibility assessment [13], but their approach lacks robust feature selection methods, neglects class imbalance considerations, and does not explore ensemble methods such as Random Forest, which have been shown to outperform traditional classifiers [14]. Collectively, existing research [9][10][11][12][13][14] exhibits critical gaps, including reliance on basic feature

selection techniques, insufficient handling of class imbalance, and a lack of comparative evaluation of ensemble methods, limiting real-world applicability gaps this research aims to bridge by integrating GA for optimized feature selection, implementing oversampling techniques, and demonstrating the superior performance of Random Forest over traditional classifiers, with future improvements focusing on region-specific data and fairness-aware AI for ethical financial decision-making [15].

3 Methods

This section provides a detailed description of the hardware and software utilized during the development process and the benchmark datasets sourced from Kaggle. It focuses on comparing the performance of three classifiers - Logistic Regression, Random Forest, and Support Vector Machine - under various feature selection techniques, namely Correlation-based selection, Recursive Feature Elimination (RFE), SelectKBest, and Lasso. Additionally, it integrates Genetic Algorithm to optimize feature selection and identifies the classifier that achieves the highest accuracy. This approach aims to enhance the predictive performance of the models and provide a robust comparison.

Fig. 1. Framework of the Study

3.1 Data Collection

For data collection, raw data will be sourced directly from Kaggle, providing a robust set of samples to support this research. Kaggle's datasets are known for their diversity, quality, and extensive real-world applications, making it an ideal repository for obtaining the data required for thorough analysis. By leveraging Kaggle's data, the study will access a well-structured sample encompassing a wide range of variables relevant to the research objectives. This will enable in-depth exploration and accurate modeling, ensuring the findings are reliable and applicable in real-world scenarios. Using Kaggle's dataset as a sample source aligns with the research's need for high-quality, authentic data that accurately represents the chosen topic

3.2 Preprocessing and Balancing

The data processing workflow begins with Data Loading, where the raw loan dataset is imported for analysis. Following this, an Exploratory Data Analysis (EDA) is conducted to gain a comprehensive understanding of the dataset's distribution, the presence of missing values, and any potential outliers that could impact the analysis or model performance. The next step is Data Cleaning, where missing values are appropriately handled, and outliers are addressed to ensure data integrity and reliability

for subsequent modeling. Finally, Data Balancing techniques, such as oversampling, are applied to the minority class (e.g., "not approved" loans) to correct class imbalances and reduce bias during model training. This sequential approach ensures the data is structured, representative, and ready for model development.

3.3 Feature Selection

A Genetic Algorithm for feature selection begins with the Initialization step, where an initial population of feature subsets is generated. Each subset is then subjected to Fitness Evaluation, where its performance is measured using a classifier's accuracy or F1-score, typically assessed through 5 K-Fold Cross-Validation to ensure robust evaluation. In the Selection phase, only the fittest individuals are chosen to progress to the next generation, ensuring that high-performing feature subsets are prioritized. These selected individuals undergo Crossover, where their features are combined to create new offspring, promoting the inheritance of beneficial feature combinations. Additionally, Mutation introduces random modifications to some features, allowing the algorithm to explore new possibilities and avoid local optima. This selection process, crossover, and mutation are repeated until a Termination criterion is met, such as reaching a maximum number of generations. Finally, the algorithm concludes with Feature Selection, where the optimal feature subset from the final generation is selected, representing the most relevant features for the classification task.

3.4 Model training and Evaluation

The model-building process involves a structured approach to classifier selection, training, and evaluation to ensure optimal performance on the loan dataset. Classifier Selection is the initial step, where suitable classifiers, such as Random Forest, Logistic Regression, and Support Vector Machine (SVM), are chosen based on the problem's requirements and the data characteristics. Each classifier offers distinct advantages: Random Forest handles non-linear relationships well, Logistic Regression is interpretable for binary outcomes, and SVM is practical for high-dimensional spaces. Model Training follows, where the selected classifiers are trained on a balanced version of the dataset, incorporating relevant features to improve the model's generalizability and predictive power. Finally, Model Evaluation is conducted using a held-out test set, employing metrics such as accuracy, precision, recall, and F1-score to gauge each model's effectiveness. These metrics provide a comprehensive assessment of model performance, enabling the identification of the most reliable classifier for predicting loan approvals.

3.5 Model Selection and Deployment

The model selection process involves identifying the best-performing model based on key performance metrics, ensuring the chosen model aligns closely with the desired outcomes and accuracy requirements. Once the optimal model is selected, it undergoes a refitting procedure where it is retrained on the entire training dataset, utilizing the most effective subset of features identified during model tuning. This refitting step ensures that the model fully leverages all available training data for improved

generalization and robustness. Finally, the refitted model is saved for future use or deployment, allowing seamless application in real-world scenarios or subsequent analyses.

4 Results & Discussion

In addressing the problem of class imbalance in a loan dataset by performing oversampling and preparing a balanced dataset for future predictive modeling. It begins with loading the dataset and conducting Exploratory Data Analysis (EDA) to understand its structure, including identifying missing values, generating summary statistics, and visualizing the original distribution of the target variable, `Loan_Status`. The EDA reveals that the dataset is imbalanced, with more instances of loan approvals ("Y") than denials ("N"), which could negatively impact the performance of machine learning models.

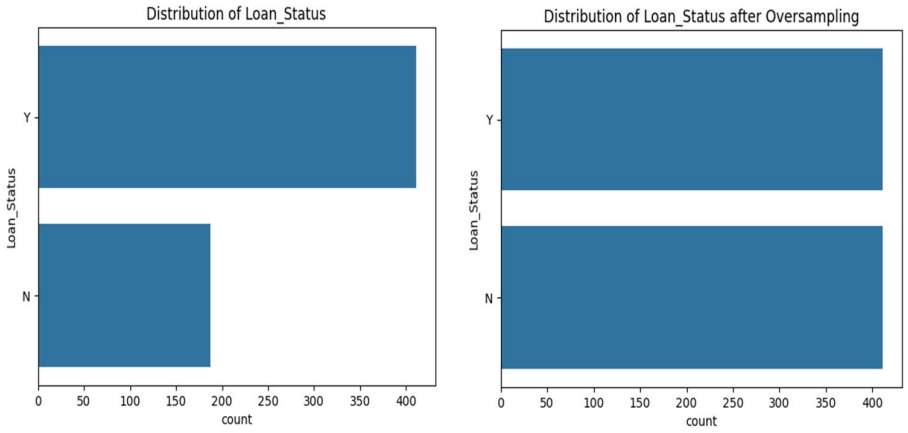

Fig. 2. Distribution of Loan Status Imbalance to Balance

To resolve this imbalance, the code implements Oversampling using the `resample` method. The minority class (`Loan_Status == 'N'`) is resampled with replacement to create a dataset where both the majority and minority classes have the same number of instances. This ensures an equal representation of both classes, which is critical for building unbiased and fair predictive models. After oversampling, the code validates the balancing process by visualizing the new distribution of `Loan_Status` and displaying the class counts.

Finally, the balanced dataset is saved as `balanced_loandata.csv`, providing a preprocessed and balanced version of the dataset for future use in machine learning experiments. This preprocessing step is crucial for improving the reliability and accuracy of predictive models that will be trained on this data.

Table 1. Accuracy Results for Different Feature Selection Methods and Classifiers

Feature Selection Method	Correlation-based	RFE	SelectKBest	Lasso
Logistic Regression	0.667	0.667	0.667	0.673
Random Forest	0.673	0.673	0.673	0.885
Support Vector Machine	0.709	0.709	0.709	0.527

In this experiment, the goal is to evaluate how different feature selection methods affect the performance of three machine learning classifiers for predicting loan approvals. The dataset used is a preprocessed and balanced version of the loan dataset, with categorical features encoded into numerical values and missing values handle. The feature selection methods applied include Correlation-Based Selection, which identifies features most correlated with the target variable; Recursive Feature Elimination (RFE), which iteratively removes the least important features; SelectKBest, which selects features based on their statistical relationship with the target variable; and Lasso Regression, which uses L1 regularization to shrink coefficients of less relevant features to zero. The classifiers utilized in this experiment are Logistic Regression, Random Forest, and Support Vector Machine (SVM). Each classifier is trained on subsets of features derived from the feature selection methods. The models are evaluated using accuracy as the primary metric, with the goal of determining the optimal combination of feature selection technique and classifier for loan approval prediction.

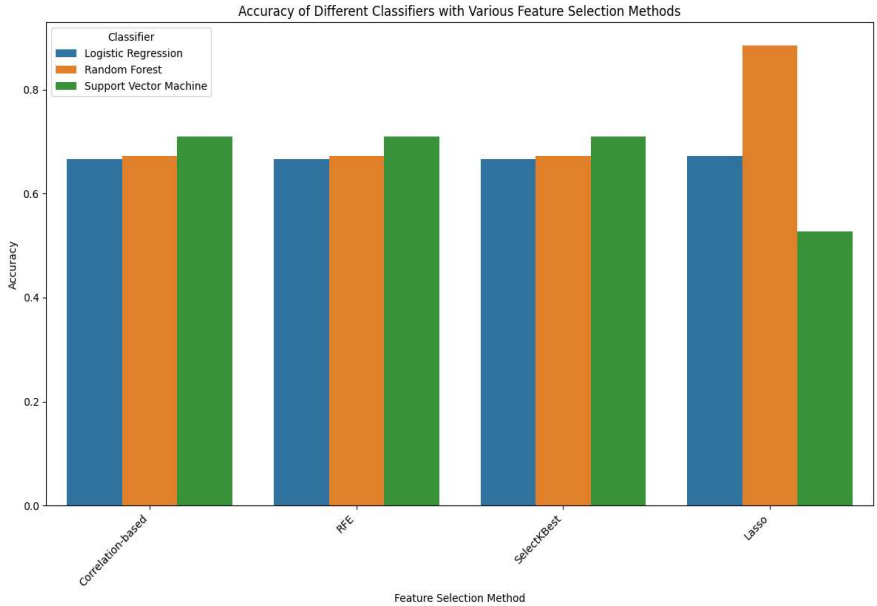

Fig. 3. Accuracy of Different Classifier with Various Feature Selection Method

The results highlight how different feature selection methods impact the accuracy of three classifiers - Logistic Regression, Random Forest, and SVM. Logistic Regression showed consistent performance across all methods, with Lasso achieving the highest accuracy (67.27%). Random Forest performed similarly with Correlation-Based, RFE, and SelectKBest (67.27%), but its accuracy significantly improved to 88.48% with

Lasso, demonstrating its ability to leverage sparse and optimized feature sets effectively. In contrast, SVM achieved its best performance (70.91%) with Correlation-Based, RFE, and SelectKBest, but its accuracy dropped significantly with Lasso (52.73%), suggesting that overly aggressive feature reduction can hinder its ability to model complex decision boundaries. Overall, Random Forest with Lasso emerged as the best-performing combination, indicating the importance of tailoring feature selection methods to the classifier's strength.

Table 2. Variables Importance Percentage

Feature	Importance	Percentage
Credit_History	0.268	26.821
ApplicationIncome	0.197	19.705
LoanAmount	0.192	19.187
CoapplicantIncome	0.105	10.544
Property_Area	0.053	5.292
Dependents	0.052	5.212
Loan_Amount_Term	0.041	4.132
Married	0.027	2.692
Self_Employed	0.023	2.320
Education	0.023	2.311
Gender	0.018	1.785

Table 2 show the evident that Credit_History is the most critical feature, contributing the approximately 28.82% to the model's predictive accuracy. This aligns with domain knowledge, where an individual's credit history often plays a significant role in determining the loan approval likelihood. Following Credit_History, ApplicantIncome (19.70%) and LoanAmount (19.19%) are the next most influential feature, collectively accounting for nearly 40% of the most importance. These features reflect an applicant's financial capability, which is a key determinant in loan decisions. Features like CoapplicantIncome (10.54%) and Property_Area (5.29%) also hold moderate importance, highlighting the relevance of combined household income and the geographical region in the decision-making process. Lower-ranked features such as Dependents, Loan_Amount_Term, and Married exhibit reduced influence, contributing between 4.13% and 2.69% each. Lastly, demographic variables like Self_Employed, Education, and Gender show minimal impact, each contributing less than 2.5% to the model's accuracy.

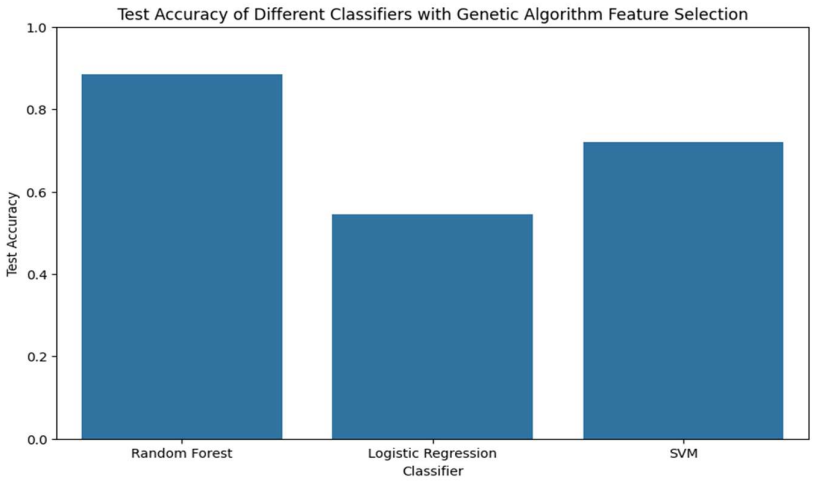

Fig. 4. Accuracy of Different Classifiers with Various Feature Selection Methods

The figure illustrates the test accuracy of three classifiers - Random Forest, Logistic Regression, and Support Vector Machine (SVM) - when applied to features selected through a Genetic Algorithm. This visualization is directly tied to the evaluation process outlined in the code, where the Genetic Algorithm optimizes feature subsets for each classifier to enhance prediction performance. It shows that Random Forest significantly outperforms the other classifiers, achieving the highest accuracy. This result reflects its ability to handle complex feature interactions effectively, which is a key strength of ensemble learning methods. In contrast, Logistic Regression demonstrates the lowest accuracy, indicating its limitations in capturing non-linear relationships within the data. This aligns with its linear nature and reliance on simpler feature interactions. SVM, while performing better than Logistic Regression, achieves moderate accuracy, reflecting its capability to handle high-dimensional data but also its sensitivity to feature scaling and kernel selection. This comparison highlights the importance of combining feature selection techniques, such as those driven by a Genetic Algorithm, with robust classifiers like Random Forest to maximize predictive performance. The findings emphasize that Random Forest is the most effective model for loan approval prediction when paired with an optimized feature subset.

Table 3. Comparative Analysis of Classifier Matrix

Feature Selection Method	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.67	0.62	0.90	0.73	0.73
SVM	0.71	0.63	0.99	0.77	0.69
Random Forest	0.85	0.86	0.84	0.85	0.94

Table 3 shows the analysis of classification models, including Logistic Regression, SVM, and Random Forest, highlights their performance across key metrics such as Accuracy, Precision, Recall, F1 Score, and AUC.

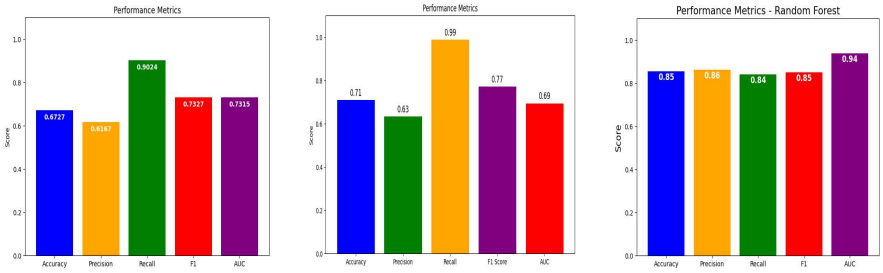

Fig. 5. Performance Matrix of 3 classifiers

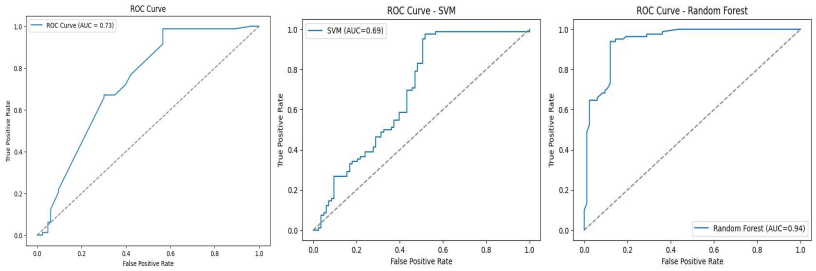

Fig. 6. ROC Curve of 3 classifiers

Logistic Regression demonstrates moderate performance with an Accuracy of 0.67 and an AUC of 0.73, with a high Recall (0.90) but low Precision (0.62), indicating a tendency to minimize false negatives. The SVM model achieves an Accuracy of 0.71 and an AUC of 0.69, with a notably high Recall (0.99) but lower Precision (0.63), suggesting strong sensitivity but potential misclassification of negatives. In contrast, the Random Forest model outperforms both, achieving an Accuracy of 0.85 and an AUC of 0.94, with well-balanced Precision (0.86) and Recall (0.84), indicating robust classification performance. The ROC curves further confirm the models' discriminative abilities, with Random Forest demonstrating the highest reliability, followed by Logistic Regression and SVM.

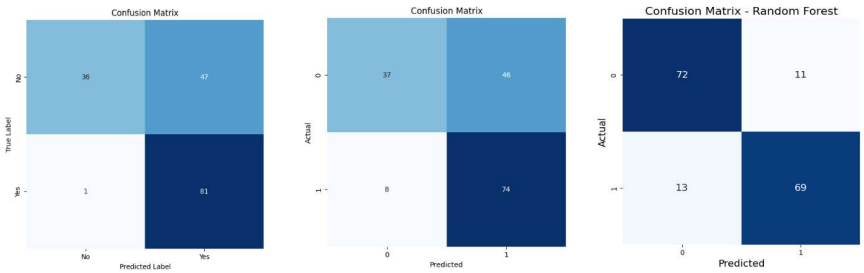

Fig 7. Confusion Matrix of 3 classifiers

The confusion matrix is used to evaluate the performance of three classifiers, Random Forest, Logistic Regression, and SVM in predicting loan approvals by categorizing predictions into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The results from the confusion matrices indicate that the Random Forest classifier provides the best balance between precision and recall, with low false positives (11) and moderate false negatives (13), making it a reliable choice for loan approval prediction. Logistic Regression performs well in approving eligible applicants (high true positives: 74) but has a high false positive rate (46), posing financial risks. SVM achieves the highest recall, with only 1 false negative, ensuring almost all eligible applicants are approved, but suffers from 47 false positives, making it unsuitable for strict financial risk management. Overall, Random Forest emerges as the most balanced and reliable classifier for loan approval prediction in this study.

Table 4. Cross-validation Accuracy Score

Feature Selection Method	1	2	3	4	5	Mean
Logistic Regression	0.764	0.703	0.677	0.707	0.683	0.707
SVM	0.752	0.697	0.701	0.738	0.726	0.723
Random Forest	0.933	0.903	0.909	0.902	0.951	0.920

The following cross validation accuracy score analyzed the performance of the three machine learning models evaluated through 5-fold cross-validation.

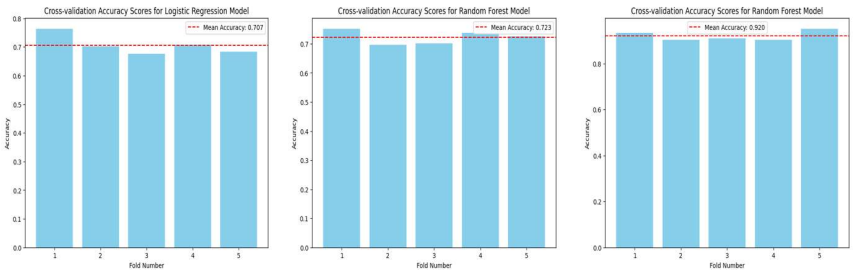

Fig. 8. 5-fold cross-validation of 3 classifiers with Genetic Algorithm

The cross-validation accuracy scores for three machine learning models Random Forest, Logistic Regression, and SVM were evaluated using a 5-fold cross-validation process. The Random Forest model demonstrated the highest performance, with accuracy scores ranging from 0.87 to 0.94 and a mean accuracy of 92.0%, indicating strong predictive capability. The Logistic Regression model showed moderate performance, with accuracy scores between 0.67 and 0.78 and a mean accuracy of 70.7%. Similarly, the SVM model exhibited accuracy scores ranging from 0.67 to 0.77, with a mean accuracy of 72.3%. Overall, Random Forest outperformed the other models, while Logistic Regression and SVM demonstrated moderate predictive power.

5 Conclusion

The study concludes that Random Forest, when integrated with Genetic Algorithm-based feature selection, is the most effective classifier for loan approval prediction. It consistently achieved the highest accuracy, precision, recall, and F1 score, demonstrating its ability to balance identifying eligible loans while minimizing false positives and false negatives. This makes it a robust and reliable model for both fairness and financial risk management. Support Vector Machine (SVM) excelled in recall, minimizing the rejection of eligible applicants, making it a suitable choice in scenarios where ensuring fairness is a priority. However, its lower precision highlighted challenges in reducing false approvals, which could increase financial risk. Logistic Regression, while interpretable and straightforward, underperformed across all metrics, indicating its limitations in capturing the complex relationships present in the dataset. The integration of Genetic Algorithms for feature selection played a critical role in optimizing model performance by identifying the most relevant features, demonstrating the importance of feature engineering in predictive modeling. Ultimately, this study identifies Random Forest as the most balanced and reliable solution, capable of addressing both operational efficiency and equitable loan decision-making processes.

References

[1] Karthiban, R., Ambika, M., & Kannammal, K. E. (2019, January). A review on machine learning classification technique for bank loan approval. In 2019 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/8822014/

[2] A. Dutta, P. (2021). A Study On Machine Learning Algorithm For Enhancement Of Loan Prediction. International Research Journal of Modernization in Engineering Technology and Science, 3. https://www.academia.edu/download/65377254/IRJMETS333632_ paper.pdf

[3] Goyal, A., & Kaur, R. (2016). A survey on ensemble model for loan prediction. International Journal of Engineering Trends and Applications (IJETA), 3(1), 32-37.

[4] Andriosopoulos, D., Doumpos, M., Pardalos, P. M., & Zopounidis, C.(2019). Computational approaches and data analytics in financial services: A literature review. Journal of the Operational Research Society,70(10), 1581-1599.https://www.tandfonline.com/doi/abs/10.1080/01605682.20 19.1595193

[5] Dimitris Andriosopoulosa , Michalis Doumposb , Panos M. Pardalosc , Constantin Zopounidis. (2018) Computational Approaches and Data Analytics in Financial Services: A Literature Review

[6] Aslam, U., Tariq Aziz, H. I., Sohail, A., & Batcha, N. K. (2019). An empirical study on loan default prediction models. Journal of Computational and Theoretical Nanoscience, 16(8), 3483- 3488. https://www.ingentaconnect.com/contentone/asp/jctn/2019/00000016/00000008/art000 65

[7] Dosalwar, S., Kinkar, K., Sannat, R., & Pise, N. (2021). Analysis of loan availability using machine learning techniques. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 9(1),15-20. https://www.researchgate.net/profile/SharayuDosalwar/publication/354367264_Analysis of_Loan_Availability_using_Machine_Learning_Techniques/links/6139b12a349f12090ff1bff6/Analysis-of-LoanAvailability-using-Machine- LearningTechniques.pdf

[8] Gupta, A., Pant, V., Kumar, S., & Bansal, P. K. (2020, December). Bank Loan Prediction System using Machine Learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE. https://ieeexplore.ieee.org/abstract/document/9336801/

[9] Ruud, M., & Nilsen, H. B. (2021). A Comparative Study in Binary Classification for Loan Eligibility Prediction (Master's thesis, Handelshøyskolen BI).

[10] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357

[11] Kiwa, F. J. Loan Eligibility System Using Machine Learning.

[12] Salter, R. I. (2023). Explainable Artificial Intelligence and its Applications in Behavioural Credit Scoring

[13] Pandey, A., & Joseph, J. (2025). Integral Role of Blockchain and Artificial Intelligence in Sustainable Economic Development. In Driving Socio-Economic Growth With AI and Blockchain (pp. 271-290). IGI Global Scientific Publishing.

[14] Chan, J. C. W., & Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment, 112(6), 2999-3011.

[15] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.