MLLoanAppPrediction: Loan Approval Prediction using Machine Learning

Diaa Salama AbdElminaam[1], Tarek Mohamed[2],Mohamed Mohamed[3]
, Ahmed Hossam Mohamed Abdelwahab[4],
George Ayman Youssry[5], Nader Amir Elhamy[6], Ramez Ehab Maurice[7]
Faculty of Computer Science
Misr International University, Cairo, Egypt
diaa.salama[1], tarek.talaat[2],mohamed.kmohamed[3],
ahmed2104740[4], george2100977[5], nader2100481 [6], ramez2100241[7]{@miuegypt.edu.eg}

***Abstract*—In this paper, we investigate the realm of loan approval prediction through the application of machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, Linear Regression, Decision Tree, and Naive Bayes. The primary objective is to assess the efficacy of these algorithms in accurately predicting loan approval outcomes based on historical data. By leveraging diverse features such as applicant information, credit history, and financial indicators, we aim to train and evaluate these algorithms to ascertain their predictive capabilities. This study contributes vital insights into enhancing loan approval systems, shedding light on the most effective machine learning techniques within the context of lending practices.**

Keywords: Loan approval; Machine Learning; Classification; Naïve Bayes; Random Forest; Decision Tree; Support Vector Machine; Linear Regression; K-Nearest Neighbor.

I. INTRODUCTION

The evaluation of loan applications poses a significant challenge for financial institutions, impacting their operational processes due to potential inaccuracies or insufficient information. Consequently, banks strive to mitigate credit risks by meticulously assessing loan statuses through extensive evaluation protocols, aiming to avert unforeseen complications. Hence, accurate prediction of loan outcomes based on provided and gathered data holds paramount importance in this context. Leveraging data mining, especially machine learning techniques, offers a promising avenue to facilitate precise and timely decisions regarding loan approvals or rejections.

Within the banking sector, loans stand as pivotal financial transactions crucial to the overall success of banks. These loans serve as the primary source of profit, forming the core assets of the banking institutions. Consequently, the foremost goal of every bank is to ensure prudent investment of its assets. Therefore, banks endeavor to mitigate credit risks by rigorously evaluating loan statuses through meticulous processes. This assessment aims to preemptively avoid any unforeseen circumstances that could impede borrowers from meeting their obligations.

Moreover, the loan approval process involves comprehensive risk assessment to safeguard the financial stability of lending institutions. The accurate estimation of potential risks associated with granting loans is pivotal, enabling banks to channel their resources toward credible and reliable borrowers while minimizing exposure to defaults or financial uncertainties. It is through this careful risk management that banks can strike a balance between profitability and prudence, ensuring sustainable growth and operational stability.

This context establishes the significance of accurate loan evaluation for financial institutions. Now, let's delve into how machine learning is revolutionizing this process. Machine learning, a subfield of artificial intelligence, focuses on developing systems that can learn and make predictions based on experiences. These systems, when applied to loan approval, aim to build models using various algorithms trained on datasets, predicting the likelihood of approving a loan based on input data encompassing diverse financial indicators.

In this research, we will focus on employing specific machine learning algorithms, including Naïve Bayes, Random Forest, Decision Tree, Support Vector Machine, Linear Regression, and K-Nearest Neighbor, to optimize loan approval processes. These algorithms, categorized under supervised learning, are chosen for their relevance and effectiveness in handling financial data, and are expected to provide valuable insights into loan approval prediction.

The integration of machine learning into the loan approval process offers unprecedented opportunities for financial institutions to streamline operations, enhance accuracy in decision-making, and minimize risks associated with granting loans. This research paper delves deeper into exploring the practical applications of these machine learning techniques in optimizing loan approval processes, evaluating their effectiveness, and providing insights into their implications for the banking sector.

The contributions made to this topic are:

- Loan Approval Prediction using machine learning.
- The testing of 6 machine learning algorithms.
- The use of 1 dataset, of 598 records, with 12 features.

The subsequent sections of this paper are organized as follows: the second section discusses related work. Additionally, the third section details the proposed research methodology, encompassing the dataset description and the employed algorithms. The fourth section presents the results derived from the

utilized algorithms and provides their analysis. The conclusion is presented in the fifth section. Furthermore, the sixth section acknowledges all the supporting figures used in this research.

II. Related Work

Predicting loan approval has been the focus of numerous studies, many of which have produced significant findings. In this paper, we used information from the literature to direct our investigation. We will properly acknowledge and quote the relevant articles that have impacted our approach to predicting loan approval outcomes in the references section.

1) Viswanatha et al. [1] addressed the challenge of accurately assessing many applicant factors. Their objective was to speed up the loan approval process by utilizing machine learning models such as K-Nearest Neighbors, Random Forest, Naive Bayes, and Decision Tree. With an accuracy of 83.73%, the Naive Bayes approach yielded the best results, however other models provided useful data.

2)Wanjun Wu's paper [2] reveals the accuracy of Random Forest and XGBoost models in loan default forecasting. The Random Forest model predicts with a high accuracy of 0.90657, while the XGBoost model achieves a higher accuracy of 0.90635. The study emphasizes the importance of feature engineering in producing reliable prediction models, highlighting their effectiveness in loan default forecasting.

3) Junhui Xu et al.'s paper [3] uses machine learning to predict loan default rates in the Chinese P2P lending industry. They used four algorithms: random forest, XGBoost, gradient boosting, and neural network models. Results showed that borrower verifications significantly impact loan default rates. The random forest model was the most accurate, with over 90% accuracy. This provides valuable guidance for risk management in the lending industry.

4) Anant Shinde et al. [4] developed a machine learning-based loan prediction system using logistic regression, decision trees, and random forests. Despite challenges like feature selection and poor data quality, the system achieved an accuracy of approximately 82%. Rating-based logistic regression performed better, highlighting the need for improved data quality.

5) Yash Divate et al.'s [5] study on loan approval prediction achieved a best-case accuracy of 0.811, emphasizing the challenges faced by applicants with lower credit scores and the importance of income levels in loan acceptance. The models used logistic regression, Dynamic K-Nearest Neighbor, and Distance and Attribute Weighted method, highlighting the need for precise forecasts to reduce non-performing assets and manage unstructured data.

6) Suliman Mohamed Fati's paper [6] proposes a machine learning model, Logistic Regression, for precise loan status approval, demonstrating its superiority over Decision Tree and Random Forest. It outperforms other methods with 91% accuracy, with an AUC of 0.8. However, the author acknowledges the need for model refinement and the lack of published outcomes in relevant literature, suggesting further research is needed to improve the model's performance.

7) Miraz Al Mamun et al. [7] developed a machine learning approach to predict bank loan eligibility using Decision Tree, Naive Bayes, and Logistic Regression. They achieved an accuracy of 92% and an F1-Score of 96%, with Gradient Boosting being the best model for predicting bank loan default. However, they faced challenges with data preparation and gathering.

8) Nitesh Pandey et al. [8] developed a machine learning model for loan approval prediction using Random Forest, Decision Tree, Support Vector Machine, and Logistic Regression methods. Logistic Regression outperforms Decision Tree in true positive and negative values. However, challenges include unbalanced datasets, feature selection, and interpretability. The Support Vector Machine model outperforms others with precision of 0.46, recall of 0.95, and F1 score of 0.61.

9) Nikhil Bansode et al.'s [9] study aims to assist financial institutions in making informed decisions about loan acceptance using machine learning algorithms. They use various methods, including K Neighbors, Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine. Logistic regression has the highest accuracy rate at 84.376%. The study emphasizes the importance of sensitivity and specificity in analyzing and displaying the ROC Curve. It also discusses the potential benefits for lenders and borrowers.

10) Shubham Singh [10] explores the application of machine learning to credit risk evaluation. It evaluates three models—Random Forest, CatBoost, and Logistic Regression—using metrics like accuracy and F1-score. Recall of 0.708 and accuracy of 68.90% indicate the good performance of the Random Forest model. CatBoost performs well, with a recall of 0.696 and accuracy of 67.96%. The Logistic Regression model achieves 65.30% accuracy and 0.661 recall. The study emphasizes the need of having high true positive forecasts and low false positives in order to identify high-risk loans.

11) Vegh et al.'s paper [11] compares 27 machine learning classification models for loan approval prediction using MATLAB and dataset analysis. The authors found that loan terms and credit scores significantly influence loan acceptance results. The models with the highest accuracy rates were neural networks and ensembles, with the optimized ensemble model achieving 98.83% precision, recall, accuracy, and

F1-score. These findings could significantly reduce loan approval processing times.

12) Hitesh K. Sharma et al. [12] present a machine learning-based model for forecasting loan amounts and distribution, aiming to improve accuracy and reduce risks in loan selection. They train supervised classification models like Random Forest, Decision Tree, and Logistic Regression using variables, bank lending guidelines, and historical loan records. The study reveals Logistic Regression as the most accurate model, with a significant accuracy score of 0.829.

13) Kasar et al.'s paper [13] on machine learning on accurately anticipating loan approvals. They found that Logistic Regression achieved 82% accuracy, 80% precision, and 85% recall, while Random Forests and Support Vector Machines performed well. This suggests that machine learning can enhance loan approval processes by reducing risks and improving speed and accuracy.

14) V. Sravan Kiran et al. [14] developed a loan eligibility prediction system using the random forest algorithm. The system gathers user data through a registration page and uses a training dataset to train and evaluate the model. The system achieves 97% train accuracy in loan eligibility prediction, surpassing the current decision tree method. The authors emphasize the random forest's accuracy in predicting loan acceptance and anticipate further advancements in accuracy and processing speed.

15) P. Bhargav et al.'s [15] study uses machine learning algorithms, specifically Logistic Regression and Random Forest, to predict loan approval default. The study divides loan applicant data into training and testing sets. The Random Forest method achieves a mean accuracy of 80.8920%, while Logistic Regression has a mean accuracy of 81.2030%. The study suggests Random Forest predicts loan acceptance more accurately than Logistic Regression.

16) Satish Jaywant Manje et al.'s [16] study uses random forest, decision tree, and logistic regression techniques to predict loan defaulters in banking. They found that logistic regression outperforms decision tree and random forest models with an accuracy of 83%, highlighting its usefulness in banking risk management.

17) Using machine learning algorithms such as Logistic Regression, Support Vector Classifier (SVC), Decision Tree, and Random Forest, Dasari et al. [17] provide a bank loan status prediction model. Recall (90% to 98%), F1-score (87% to 91%), precision (82% to 99%), and accuracy (94%) are all markedly improved by the model. The model is a viable strategy for improving loan approval procedures and accuracy in the banking industry, and the paper makes recommendations for further improvements, including adding new features and using neural network frameworks like PyTorch and Tensorflow.

18) K. Malathi et al. [18] examine a machine learning strategy for loan approval prediction, comparing Random Forest and Decision Tree algorithms. They found that the Random Forest method outperforms Decision Tree with an accuracy of 79.4490%. The study provides a detailed description of the algorithms, their functionality, and statistical analysis, highlighting the potential of machine learning algorithms in loan approval prediction.

19) Pallapothu Nishita et al. [19] conducted a study on loan approval prediction using machine learning methods on a large dataset. The models showed accuracy ranging from 75% to 97%, with Logistic Regression being the most accurate at 88.70%. The study highlighted the importance of heterogeneous attributes in improving loan approval prediction models.

20) In order to forecast loan eligibility in banking, Ugochukwu.E. Orji et al. [20] investigate machine learning algorithms. They use six different algorithms on a historical dataset, and they achieve great accuracy 95% with the random forest model, for example. Robust prediction models are produced by ensemble approaches and methodologies such as SMOTE, which demonstrate potential enhancements for loan approval procedures in the financial industry.

III. Proposed Methodology

Numerous algorithms were used, and a research was done on each algorithm before training the model using them on the datasets. The following diagram represents the steps the datasets went through to get the results.

A. Datasets Descriptions

The dataset [21] consists of 12 features, with 598 records. The feature of Loan ID was removed as it does not affect the results in any way. The detailed features of this dataset can be viewed in Table I and Table II

TABLE I
FEATURES' DESCRIPTION OF DATASET 1

Feature	Description
Gender	Gender of the applicant
Married	Marital Status of the applicant
Dependents	Number of dependants if any
Education	Educational Status
Self Employed	Defines if the applicant is self employed
Applicant Income	Applicant income
Coapplicant Income	Co-applicant income
Loan Amount	Loan amount (in thousands)
Loan Amount Term	Terms of loan (in months)
Credit History	Credit history of individual's repayment of debts
Property Area	Area of property
Loan Status (Target)	The acceptance or rejection of the loan

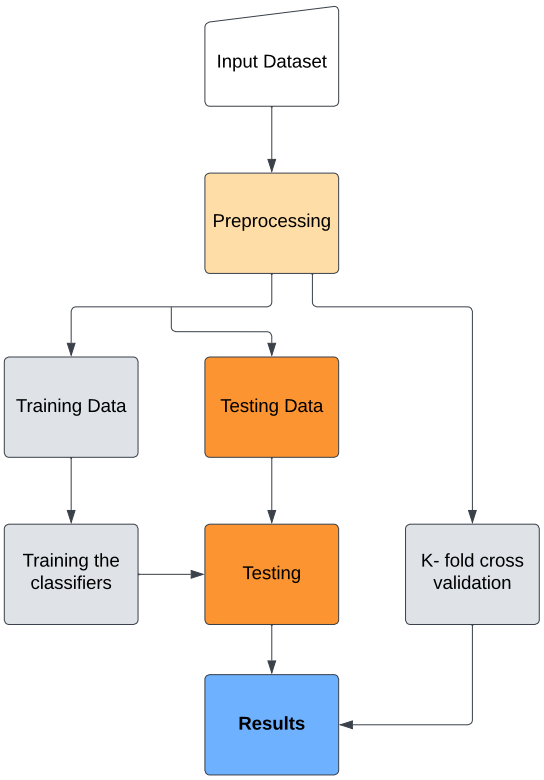

Fig. 1. Methodology Overview

TABLE II
FEATURES' DETAILS OF DATASET 1

Feature	Type	Values
Gender	Categorical	Male or Female
Married	Categorical	Yes or No
Dependents	Categorical	0, 1, 2 or 3
Education	Categorical	Graduate or Not Graduate
Self Employed	Categorical	Yes or No
Applicant Income	Numerical	From 150 to 81000
Coapplicant Income	Numerical	From 0 to 41667
Loan Amount	Numerical	From 9 to 650
Loan Amount Term	Numerical	From 12 to 480
Credit History	Categorical	0 or 1
Property Area	Categorical	Urban, Semiurban or Rural
Loan Status (Target)	Categorical	Yes or No

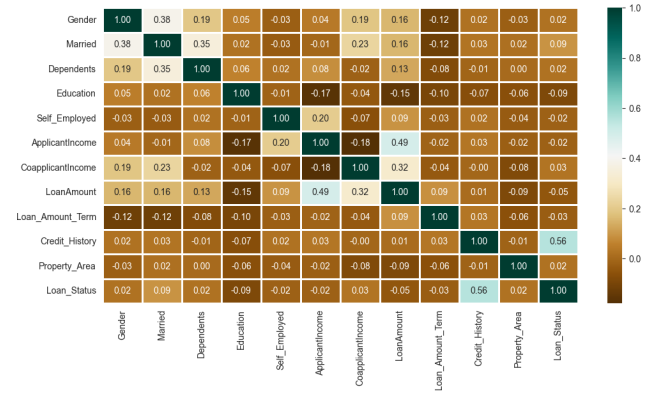

Fig. 2. Features correlations

It can be seen that Credit History has the most effect on the target.

B. *Preprocessing*

Before passing the data to the machine learning models, it has to go through some preprocessing to ensure the models will receive valid data that is ready to be processed. This includes dealing with any missing or NULL values, normalizing the data and detecting outliers.

1) Imputing Missing Data: Imputing involves filling missing values in a dataset, ensuring completeness for analysis and model training. Imputation prevents data loss, maintains statistical properties, enhancing machine learning model performance. It addresses bias and enables models to learn effectively from available information.

The number of occurrences of missing data for each feature is as follows: out of the 12 features, only 4 features had missing values which were dependants, loan amount, loan amount term and credit history.

TABLE III
COUNT OF MISSING VALUES PER FEATURE

Feature	Missing Values Count
Dependents	12
Loan Amount	20
Loan Amount Term	14
Credit History	45

For the dependants and credit history, they are categorical features, so the missing values were replaced with the feature's mode (most repeated value). The loan amount and loan amount term are numerical values, so the missing values were replaced with the feature's mean (average) value.

2) Normalization: Normalization of data ensures scale consistency among the different features of the dataset, preventing any single feature from dominating the learning process due to differences in scale, especially if the algorithm used is scale sensitive, such as kNN or SVM.

The method used is Z-score normalization, also known as standardization. It transforms the dataset into a standard normal distribution, ensuring that features have similar scales. The Z-score for each data point in a feature is calculated using the formula:

$$Z_i = \frac{x_i - \mu}{\sigma}$$

Here, Z_i represents the Z-score for the data point x_i, μ is the mean of the feature, and σ is the standard deviation of the feature. The resulting Z-scores have a mean of 0 and a standard deviation of 1, centering the values around zero and expressing them in terms of standard deviations from the mean, so all the values are now expressed as values between 0 and 1.

3) Detecting Outliers: Outliers are considered anomalies that can distort statistical measures and impact model performance. Detecting them is vital in the preprocessing phase of the data. By removing these anomalies, more accurate insights and predictions can be ensured, preventing skewed results.

To detect the outliers in this dataset, the Z-score method analysis is utilized. The Z-scores of each data point is calculated, and a threshold is specificed. If the Z-score exceeds the threshold in the positive or negative direction, the data is flagged as an outlier and excluded from the dataset. Using a threshold of 3.5, the outliers count in each feature is as follows:

TABLE IV
OUTLIERS' COUNT FOR THE DATASET

Feature	Outliers Count
Applicant Income	6
Coapplicant Income	4
Loan Amount	12
Loan Amount Term	9

All the records containing any outlier data is excluded from the dataset. Out of the original 598 records, 569 are left.

4) Resampling: The data is divided into training and testing data. After removing the outliers and checking the distribution of the dataset, it is seen that it is unbalanced, with almost 70% of the data having 'Y' as a target and less than 30% with a target of 'N'.

To avoid overfitting the models, class imbalance will be addresed by oversampling the minority class 'N'. The oversampling technique used is SMOTE (Synthetic Minority Oversampling Technique). The training data was passed over to SMOTE and the results were displayed in a piechart as shown in Fig. 4.

Now the training set contains almost equal proportion of 'Y' target as 'N' target. The data is now ready to be passed to the models.

C. Used Algorithms

1) Random Forest: It is a regression and classification technique for group learning. During training, it builds a large

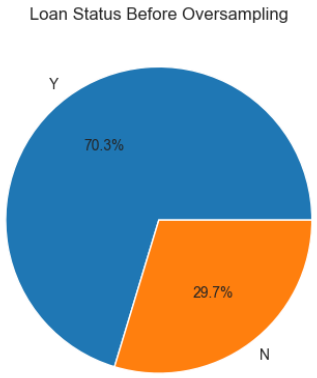

Fig. 3. Target distribution

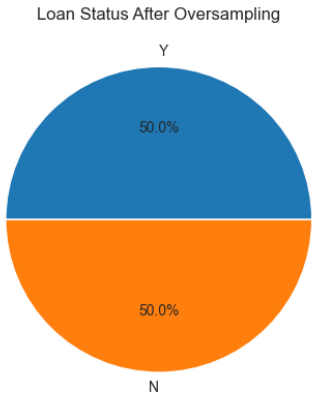

Fig. 4. Target distribution after SMOTE

number of decision trees, and at output, it outputs the class that is the mean prediction (regression) or the mode of the classes (classification) of the individual trees.

The tendency of decision trees to overfit to their training set is compensated for by random decision forests.

Although they are less accurate than gradient enahnced trees, random forests still perform better than choice trees in most cases. Their performance, however, may be impacted by the peculiarities of the data.

2) K - Nearest Neighbours: The k-nearest neighbours algorithm (k-NN) is a non-parametric technique for regression and classification in pattern recognition. The k closest training instances in the feature space make up the input in both scenarios. Whether k-NN is used for regression or classification determines the outcome.

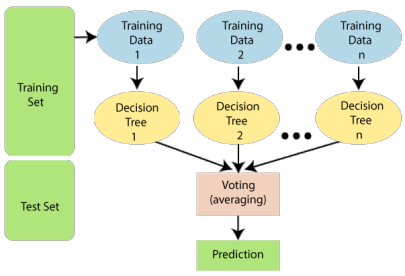

Fig. 5. Flowchart of Random Forest Algorithm [22]

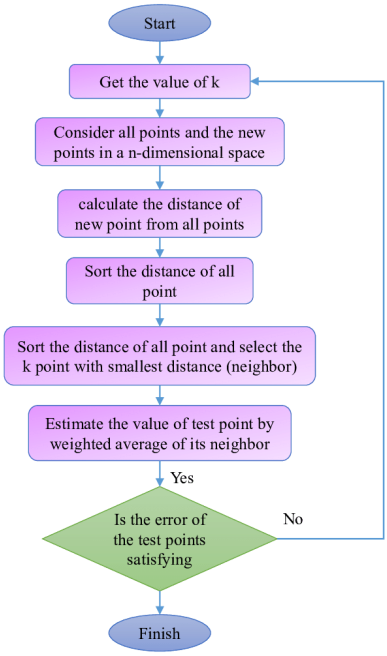

Fig. 6. Flowchart of K-Nearest Neighbours [23]

3) Support Vector Machine: Support-vector machines, also known as support-vector networks, are supervised learning models that examine data used in regression and classification studies. They are paired with learning algorithms.

An SVM training algorithm creates a model that assigns new examples to one or the other of two categories given a set of training examples that are each marked as belonging to one or the other of two categories; this makes the model a non-probabilistic binary linear classifier (although Platt scaling can be used to use SVM in a probabilistic classification setting).

4) Naïve Bayes: Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorm with strong (naive) assumptions between the features

Formula: $P(A|B) = \frac{P(B|A)*P(A)}{P(B}$

5) Logistic Regression: For classification and predictive analytics, this kind of statistical model—also referred to to

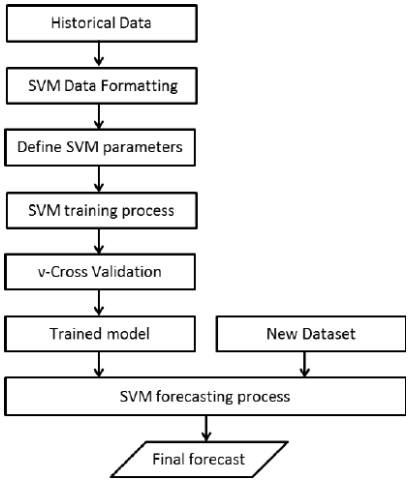

Fig. 7. Flowchart of Support Vector Machine [24]

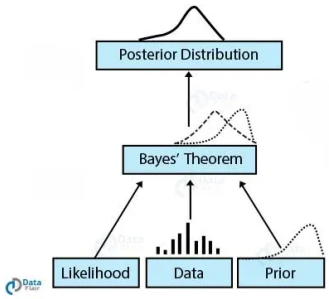

Fig. 8. Flowchart of Naïve Bayes [22]

as the logit model—is widely employed. Logistic regression uses a dataset of random variables to estimate the likelihood of an event occurring, such as voting or not. Because the result is a probability, the dependent variable has a range of 0 to 1. A logit transformation is performed to the odds in logistic regression, which is the probability of success divided by the probability of failure.

Formula: $P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

6) Decision Tree: Using a decision tree as a predictive model, decision tree learning goes from observations about an object (shown as branches) to inferences about the item's target value (shown as leaves). It is a predictive modeling technique applied to data mining, machine learning, and statistics. Classification trees are tree models in which the target variable is discretely variable; leaves in these tree structures represent class labels, and branches represent conjunctions of features that lead to those class labels.

Regression trees are decision trees in which the goal variable is capable of taking continuous values, usually real numbers. A decision tree can be used in decision analysis to formally and visually reflect decisions and decision-making.

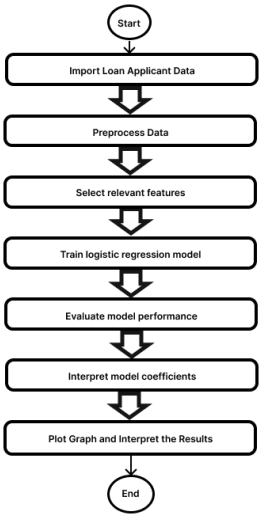

Fig. 9. Flowchart of Logistic Regression

In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making)..

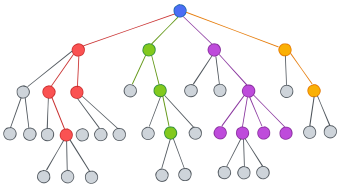

Fig. 10. Decision Tree Visualization

D. *Performance Metrics*

The performance for each classifier used was measured against four main metrics. For the following metrics these terms are used:

- TP stands for True Positives, referring to the number of instances where the model correctly identified the positive class.
- TN stands for True Negatives, referring to the number of instances where the model correctly identified the negative class.
- FP stands for False Positives, referring to the number of instances where the model predicted a positive class, but the true class was negative.
- FN stands for False Negatives, referring to the number of instances where the model predicted a negative class, but the true class was positive.

1) Accuracy: It is the number of correct predictions divided by the total number of predictions.

It is calculated as $\frac{TP + TN}{TP + TN + FP + FN}$.

2) Precision: It is the ratio of true positive predictions to the total predicted positives.

It is calculated as $\frac{TP}{TP + FP}$.

3) Recall: It measures the ability of the classifier to capture all positive instances.

It is calculated as $\frac{TP}{TP + FN}$.

4) F1 Score: It is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance, particularly in situations where precision and recall may be in conflict.

It is calculated as $2 \times \frac{Precision \times Recall}{Precision + Recall}$.

IV. Results and Analysis

After training and testing the models, the results collected from Random Forest, kNN, SVM, Naive Bayes, Logistic Regression and Decision Tree are shown below.

The following results are from the dataset with a 80/20 data split.

TABLE V
STATISTICS OF CLASSIFIERS WITH 80/20 DATA SPLIT

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.807	0.783	0.973	0.867
Naïve Bayes	0.798	0.780	0.959	0.861
Logistic Regression	0.772	0.786	0.892	0.835
Decision Tree	0.754	0.838	0.770	0.803
Random Forest	0.746	0.800	0.811	0.805
kNN	0.640	0.746	0.676	0.709

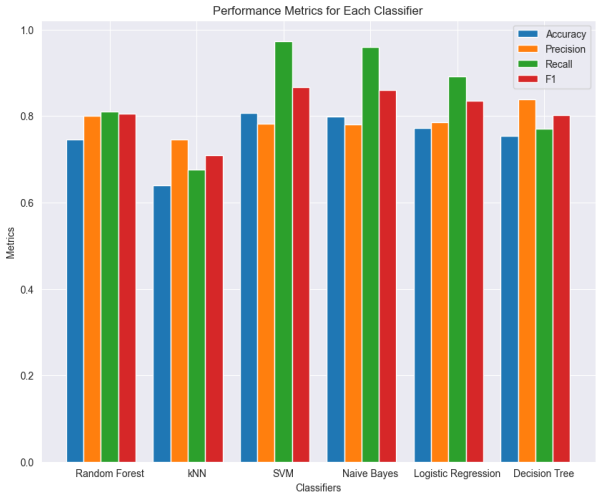

Fig. 11. Performance with 80/20 data split

SVM was the most accurate algorithm with an accuracy of 0.807 and F1 score of 0.861, closely followed by Naive Bayes

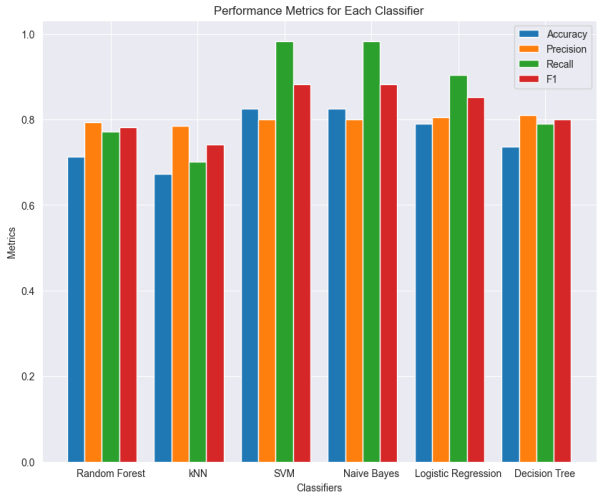

Fig. 12. Performance with 70/30 data split

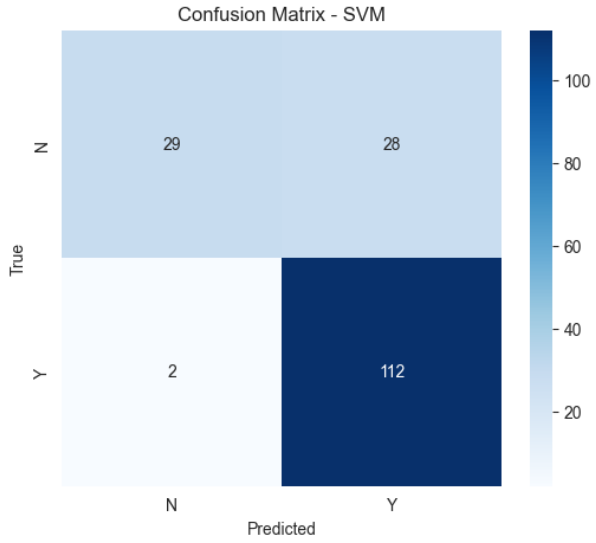

Fig. 13. SVM Confusion Matrix with 70/30 split

with an accuracy of 0.798. Decision Tree had the highest precision of 0.838 while kNN had the worst performance across the 4 metrics, with an accuracy of 0.640.

Now the results are recorded again but with a 70/30 data split.

TABLE VI
STATISTICS OF CLASSIFIERS WITH 70/30 DATA SPLIT

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.825	0.800	0.982	0.882
Naïve Bayes	0.825	0.800	0.982	0.882
Logistic Regression	0.789	0.805	0.904	0.851
Decision Tree	0.737	0.811	0.789	0.800
Random Forest	0.713	0.793	0.772	0.782
kNN	0.673	0.784	0.702	0.741

SVM and Naive Bayes came in these results as the most performing classifiers sharing equal statistics against the 4 metrics used. They had the highest accuracy of 0.825 and the highest F1 score of 0.882, which is a slight improvement from the previous results. Decision Tree remained the algorithm with the highest precision of 0.811, a slight decrease from the previous results, and kNN still came last in terms of all metrics, with an accuracy of 0.673.

TABLE VII
STATISTICS OF CLASSIFIERS WITH 10K FOLD

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.806	0.766	0.881	0.819
kNN	0.756	0.749	0.772	0.760
Decision Tree	0.745	0.742	0.751	0.747
Naïve Bayes	0.722	0.646	0.985	0.780
Logistic Regression	0.721	0.649	0.962	0.775
SVM	0.721	0.645	0.985	0.779

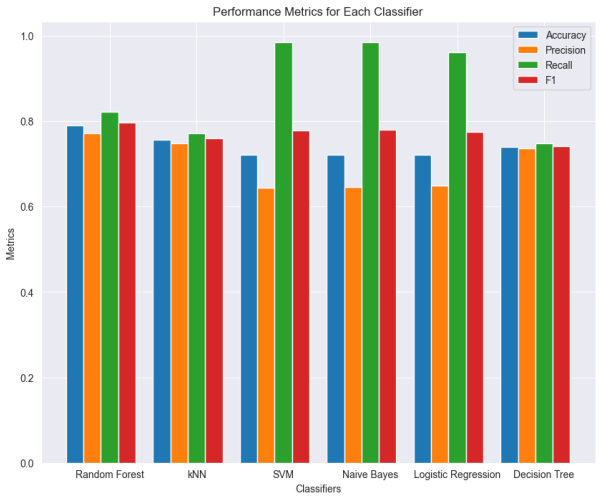

Fig. 14. Performance with 10k fold

Using k-fold on the dataset had slight improvements with some algorithms and noticeable decline in others. Random Forest had the highest accuracy of 0.806, the highest Precision of 0.766 and the highest F1 score of 0.819. KNN had an accuracy of 0.756, which was a great increase in accuracy compared to the initial results. Both SVM and Naive Bayes remained the highest in terms of Recall with a value of 0.985,

but had a noticeable decline in the other 3 metrics. Logistic Regression and Decision Tree both suffered some degredation as well.

V. Conclusion

While loan requests keep increasing as economies keep rising, banks and insurance companies are guaranteed to have loads of applications that would need proper, fast and reliable reviewing to study each case. With the help of machine learning, this can be done almost instantly by letting the models predict if the person is applicable for a loan or not. From the results, SVM, Naive Bayes and Random Forest proved to give great results with accuracies of 82.5%, 82.5% and 80.6%.

While machine learning can be used in any field, it is very helpful in fields where there are many details or records to be reviewed in short periods of time, such as this one. The performance of the models can be improved even further using local banks' or companies' loan applications' data as input datasets, training the model on specific data for local banks rather than using a public dataset.

VI. Acknowledgment

First and foremost, we would like to show our gratitude for the staff working in Misr International University and the faculty of computer science for working hard to make this university a success. We are very thankful for Prof. Mohamed Shebl El Komy University President, Prof. Ayman Bahaa dean faculty of Computer Science, Prof. Abdelnasser Zaied Vice Dean of Student Affairs and Professor of Computer Engineering, and Dr. Ayman Nabil Associate Professor in Computer Science for giving us the opportunity to learn in this virtuous university and running it flawlessly. And finally, we would like to give special thanks to Dr. Diaa AbdelMoneim associate professor in information systems and Eng. Mostafa Radwan teaching assistant for their continued guidance and support for our work.

References

[1] V. Viswanatha, A. Ramachandra, K. N. Vishwas, and G. Adithya, "Prediction of loan approval in banks using machine learning approach," *International Journal of Engineering and Management Research*, vol. 13, no. 4, August 2023, available at SSRN: https://ssrn.com/abstract=4532468.

[2] W. Wu, "Machine learning approaches to predict loan default," *Intelligent Information Management*, vol. 14, no. 5, pp. 157–163, 2022.

[3] Z. L. Junhui Xu and Y. Xie, "Loan default prediction in chinese p2p lending using machine learning," *Scientific Reports*, vol. 11, no. 1, p. 18759, 2021.

[4] Shinde, Anant, Patil, Yash, Kotian, Ishan, Shinde, Abhinav, and Gulwani, Reshma, "Loan prediction system using machine learning," *ITM Web Conf.*, vol. 44, p. 03019, 2022. [Online]. Available: https://doi.org/10.1051/itmconf/20224403019

[5] Y. Divate, P. Rana, and P. Chavan, "Loan approval prediction using machine learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 08, 2021. [Online]. Available: https://www.irjet.net/archives/V8/i5/IRJET-V8I5331.pdf

[6] S. Mohamed Fati, "Machine learning-based prediction model for loan status approval," *Journal of Novel Undergraduate Science*, vol. 48, no. 10, 2021. [Online]. Available: http://jonuns.com/index.php/journal/article/view/783

[7] M. Al Mamun, A. Farjana, and M. Mamun, "Predicting bank loan eligibility using machine learning models and comparison analysis," *7th North American International Conference on Industrial Engineering and Operations Management*, June 11 2022, https://doi.org/10.46254/NA07.20220328.

[8] N. Pandey, R. Gupta, S. Uniyal, and V. Kumar, "Loan approval prediction using machine learning algorithms approach," *International Journal of Innovative Research in Technology*, vol. 8, no. 1, p. 898, 2021. [Online]. Available: https://ijirt.org/master/publishedpaper/IJIRT151769_PAPER.pdf

[9] N. Bansode, A. Verma, A. Sharma, and V. Bhole, "Predicting loan approval using machine learning," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 04, no. 05, p. 375, May 2022, impact Factor: 6.752. [Online]. Available: https://www.irjmets.com/uploadedfiles/paper//issue_5_may_2022/22109/final/fin_irjmets1652288922.pdf

[10] S. Singh, "Application of machine learning on loan risk analysis," *Technical Report*, 2023.

[11] L. V'egh, K. Czak'oov'a, and O. Tak'ač, "Comparing machine learning classification models on a loan approval prediction dataset," *International Journal of Advanced Natural Sciences and Engineering Researches*, vol. XX, pp. XX–XX, 2023.

[12] H. K. Sharma, T. Choudhury, P. Ahlawat, S. N. Mohanty, and S. Jain, "Machine learning based model for loan amount prediction and distribution," *CEUR Workshop Proceedings*, vol. 3283, p. 100, 2023. [Online]. Available: https://ceur-ws.org/Vol-3283/Paper100.pdf

[13] M. S. Kasar, V. P. Mulik, and A. L. Yadav, "Modern loan approval prediction system based on machine learning," *Journal For Basic Sciences*, vol. 23, pp. 365–371, 2023.

[14] V. S. Kiran, B. T. Reddy, D. U. Kumar, K. S. A. Varma, and T. S. Kiran, "Loan eligibility prediction using machine learning," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 11, no. VIII, Aug 2023, ©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894. [Online]. Available: www.ijraset.com

[15] P. Bhargav and K. Malathi, "Using machine learning, the random forest algorithm and logistic regression to predict default loan approval," *Journal of Survey in Fisheries Sciences*, vol. 10, no. 1S, pp. 1814–1824,

2023. [Online]. Available: https://sifisheriessciences.com/journal/index.php/journal/article/download/415/398/793

[16] P. S. J. Manje, M. D. R. Manje, M. R. P. Bhere, and M. R. K. Pawade, "Loan prediction model using logistic regression, decision tree, random forest," *International Journal for Research in Engineering Application & Management (IJREAM)*, vol. 07, no. Special Issue, May 2021. [Online]. Available: https://ijream.org/papers/IJREAMV07I02SJ009.pdf

[17] Y. Dasari, K. Rishitha, and O. Gandhi, "Prediction of bank loan status using machine learning algorithms," *International Journal of Computing and Digital Systems*, vol. 14, no. 01, p. 13, 2023, received 23 Jun. 2022, Revised 6 May. 2023, Accepted 8 May. 2023, Published 1 Jul. 2023. [Online]. Available: https://journal.uob.edu.bh/bitstream/handle/123456789/4858/IJCDS140113.pdf?sequence=3&isAllowed=y

[18] P. Bhargav and K. Malathi, "Using machine learning, the random forest algorithm and logistic regression to predict default loan approval," *Journal of Survey in Fisheries Sciences*, vol. 10, no. 1S, pp. 1814–1824, 2023. [Online]. Available: https://sifisheriessciences.com/journal/index.php/journal/article/download/415/398/793

[19] P. Nishita, B. Bhowmik, P. R. Gayani, A. R, and S. P. P. Singh, "Loan approval prediction," *International Journal of Advances in Engineering & Management*, 2023, date of Submission: 05-04-2023, Date of Acceptance: 15-04-2023. [Online]. Available: https://ijaem.net/issue_dcp/Loan%20Approval%20Prediction.pdf

[20] U. Orji, C. Ugwuishiwu, J. Nguemaleu, and Ugwuanyi, "Machine learning models for predicting bank loan eligibility," 06 2022.

[21] A. Parjapat. (2017) Loan prediction. [Online]. Available: https://www.kaggle.com/datasets/ninzaami/loan-predication

[22] Javatpoint, "Example image from javatpoint [online]," image accessed from the Javatpoint website. [Online]. Available: https://www.javatpoint.com/example/page/

[23] A. Hemmati-Sarapardeh. (2020) A simple flowchart for the k-nearest neighbor modeling. ResearchGate. Accessed 16 Dec, 2023. [Online]. Available: https://www.researchgate.net/figure/A-simple-flowchart-for-the-k-nearest-neighbor-modeling_fig1_346429285

[24] L. Cipcigan. (2013) Operation flow chart of the svm model. ResearchGate. Accessed 16 Dec, 2023. [Online]. Available: https://www.researchgate.net/figure/Operation-Flow-Chart-of-the-SVM-Model_fig1_261040572