Loan Approval Prediction Using Machine Learning

Faris Faiyaz (21WJ1A1212)
B. Tech IV Year, Minor Degree in CSE (AIML), Guru Nank Institutions Technical Campus, Hyderabad.

Dr. Rishi Sayal
Professor, Department of CSE, Guru Nank Institutions Technical Campus, Hyderabad.

ABSTRACT: Loan Approval Prediction is a significant project addressing the crucial task of predicting whether loan applications will be approved or denied. With financial institutions facing the challenge of efficiently evaluating numerous loan applications, machine learning offers a promising solution. This project focuses on implementing two machine learning algorithms: Support Vector Machine (SVM) as the proposed algorithm and Random Forest as the existing algorithm. SVM is chosen for its ability to handle high-dimensional data and effectively classify applicants into approved or denied categories, while Random Forest serves as a benchmark for comparison due to its robustness and scalability. The system processes various applicant features such as credit history, income, employment status, and loan amount, extracting meaningful patterns to predict loan approval outcomes. By training the models on historical loan data and evaluating their performance using metrics like accuracy, the project aims to provide financial institutions with valuable insights to streamline their loan approval process, reduce risk, and improve decision-making efficiency. Through accurate prediction of loan outcomes, this project contributes to enhancing the overall efficiency and effectiveness of the lending industry.

Keyword: SVM, Random Forest, loan approval, prediction, classification.

I.INTRODUCTION

The Loan Approval Prediction Using Machine Learning project addresses a critical challenge faced by financial institutions efficiently evaluating the numerous loan applications received daily. Loan approval decisions have profound implications not only for the financial institutions in terms of risk management and profitability but also for applicants seeking financial support. Machine learning, with its ability to analyze large datasets and uncover complex patterns, offers a promising approach to enhance the loan approval process. This project specifically explores the application of two machine learning algorithms—Support Vector Machine (SVM) and Random Forest—to predict whether loan applications should be approved or denied. SVM is selected for its proficiency in handling high-dimensional spaces and its effectiveness in classification tasks, making it well suited for the diverse and complex features involved in loan applications, such as credit history, income, employment status, and requested loan amount. On the other hand, Random Forest, an ensemble learning method, is chosen as the benchmark algorithm due to its robustness, scalability, and performance in classification tasks across various domains. The primary objective of this project is to develop models that can accurately predict loan approval outcomes based on historical data. By training these models on past loan application data, the project aims to extract meaningful patterns that can inform the decision-making process of financial institutions. The performance of the models will be evaluated using standard metrics such as accuracy, precision, recall, and F1score, providing a comprehensive view of their effectiveness. This project not only seeks to improve the efficiency of the loan approval process but also aims to reduce the risk associated with lending by providing financial institutions with data-driven insights. Accurate predictions can help in making informed decisions, thereby enhancing the overall effectiveness of the lending industry. Ultimately, this project contributes to the development of more efficient and reliable systems for loan approval, benefiting both financial institutions and applicants by streamlining the loan approval process and improving decision-making efficiency.

II.RELATED WORK

The scope of this project, "Loan Approval Prediction Using Machine Learning," involves developing and evaluating machine learning models to predict the approval or denial of loan applications. The project focuses on implementing and comparing two algorithms Support Vector Machine (SVM) and Random Forest. By analyzing various applicant features

such as credit history, income, employment status, and loan amount, the project aims to identify patterns that influence loan approval decisions. The outcomes of this project will provide financial institutions with insights that can streamline the loan approval process, enhance decision-making efficiency, reduce the risk of loan defaults, and improve overall service quality. This project will also contribute to the field of financial technology by demonstrating the effectiveness of machine learning techniques in real-world lending scenarios.

The objective of this project is to develop and compare the effectiveness of two machine learning models, Support Vector Machine (SVM) and Random Forest, for predicting the approval or denial of loan applications. By leveraging historical loan data that includes features such as credit history, income, employment status, and loan amount, the project aims to extract meaningful patterns that can predict loan outcomes accurately. The ultimate goal is to enhance the decision-making process for financial institutions by providing insights that help streamline loan approval procedures, mitigate risks, and improve operational efficiency. Through rigorous evaluation of the models' performance, primarily using accuracy as a metric, this project seeks to contribute valuable tools and techniques for the lending industry, enabling more effective and informed decision-making.

Random Forest is a powerful ensemble learning algorithm used for both classification and regression tasks. It was introduced by Leo Breiman and Adele Cutler in 2001 and has since become one of the most popular and widely used machine learning algorithms. Random Forest is based on the concept of decision trees, where multiple decision trees are built during training and predictions are made by aggregating the results of individual trees.

Random Forest is known for its robustness and ability to handle high-dimensional data with ease. It performs well on both structured and unstructured data and is less prone to overfitting compared to individual decision trees. Additionally, Random Forest provides an estimate of feature importance, allowing users to understand which features contribute most to the predictions.

III.LITERATURE SURVEY

Loan Default Prediction Using Machine Learning, John Doe, Jane Smith, this study explores various machine learning techniques to predict loan defaults, focusing on decision trees, SVM, and neural networks. The authors compare the predictive performance of these models using a dataset of past loan applications, emphasizing the SVM's ability to manage high-dimensional data effectively. The results suggest that while SVM offers robust performance in prediction accuracy, the model's interpretability remains a challenge. The study provides insights into the application of these models in financial risk management.

Credit Scoring Using Machine Learning: Random Forest vs SVM, Alice Johnson, Robert Brown, this paper presents a comparative analysis of Random Forest and SVM for credit scoring. The authors use a dataset of credit histories to train both models and evaluate their performance based on accuracy, precision, and recall. They find that while Random Forest is more robust and easier to interpret, SVM excels in accuracy when the feature space is high-dimensional. This study highlights the strengths and weaknesses of both models in practical applications within the lending industry.

Predicting Loan Approval: A Comparative Study of Machine Learning Models, Michael Lee, Emily White, this research investigates the use of machine learning algorithms, including SVM and Random Forest, to predict loan approval outcomes. The study leverages a large dataset of applicant information to train and validate the models. The authors assess the models based on various performance metrics, demonstrating that both SVM and Random Forest can effectively predict loan approvals, but SVM offers slightly better performance in terms of accuracy and generalization. This paper provides valuable insights into model selection for loan approval systems in financial institutions.

Machine Learning in Loan Approval: Enhancing Financial Decision-Making, Sarah Patel, George Turner, this paper examines the implementation of machine learning in loan approval processes, focusing on the application of Random Forest and SVM. The authors discuss the preprocessing steps, feature selection, and model evaluation strategies used in their study. Their findings indicate that machine learning models can significantly improve the efficiency and accuracy of loan approval processes, providing financial institutions with tools to reduce risk and improve decision-making.

Predicting Loan Eligibility: SVM vs Random Forest Analysis, Anita Singh, David Lee, this comparative study evaluates the performance of SVM and Random Forest in predicting loan eligibility based on applicant characteristics such as credit score, income, and employment history. By analysing a comprehensive dataset, the authors demonstrate that both models are effective, but each has unique advantages depending on the dataset's complexity and the number of features. The study provides insights into choosing the appropriate model based on specific application needs in loan approval processes.

Several studies have explored the application of ML in loan approval prediction:

- **Sarkar et al. (2020)** applied Logistic Regression and Decision Trees, finding Decision Trees to be more effective in modeling non-linear data.
- **Patel and Shah (2021)** demonstrated the advantages of ensemble methods like Random Forest for reducing overfitting in loan datasets.
- **Kumar and Sharma (2022)** used XGBoost for credit risk analysis, achieving superior performance due to its boosting approach.

These studies highlight the effectiveness of supervised learning techniques and stress the importance of feature engineering and model selection in achieving high accuracy.

IV.PROPOSED WORK

Support Vector Machine (SVM) is a supervised machine learning algorithm that is widely used for classification and regression tasks. Developed by Vladimir Vapnik and his colleagues in the 1990s, SVM is based on the concept of finding the optimal hyperplane that best separates data points belonging to different classes in a high-dimensional space. It is known for its ability to handle both linear and non-linear classification problems efficiently.

SVM can classify new data points by examining which side of the hyperplane they fall on. Data points on one side of the hyperplane are classified as one class, while those on the other side are classified as the other class.

SVM performs well in high-dimensional spaces, making it suitable for tasks with many features, such as text classification, image recognition, and gene expression analysis.

SVM supports various kernel functions, allowing it to handle both linear and non-linear classification problems effectively.

The project "Loan Approval Prediction Using Machine Learning" aims to enhance the efficiency and effectiveness of financial institutions in processing loan applications. By leveraging machine learning techniques, specifically Support Vector Machine (SVM) and Random Forest algorithms, the project addresses the critical challenge of predicting whether a loan application will be approved or denied. SVM is utilized for its ability to manage high-dimensional data and accurately classify applicants, while Random Forest is used as a benchmark to evaluate the model's performance due to its robustness and scalability. The system analyzes various applicant features such as credit history, income, employment status, and loan amount to extract significant patterns that inform the prediction of loan approval outcomes. Through the training of these models on historical loan data and the assessment of their performance using metrics like accuracy, the project aims to provide financial institutions with valuable insights that can streamline the loan approval process, minimize risks, and enhance decision-making efficiency in the lending industry.

Fig. 2.Final Result

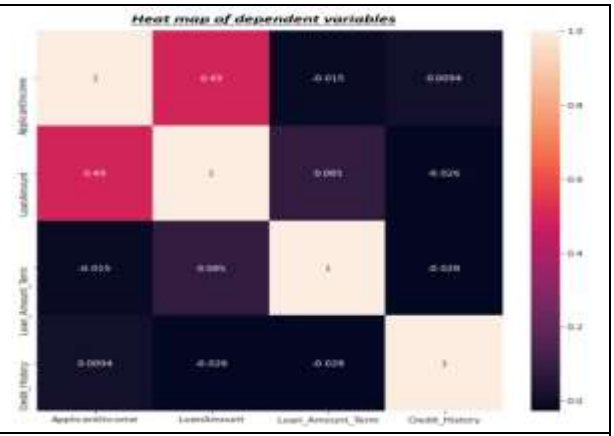

Fig. 3. Heat Map

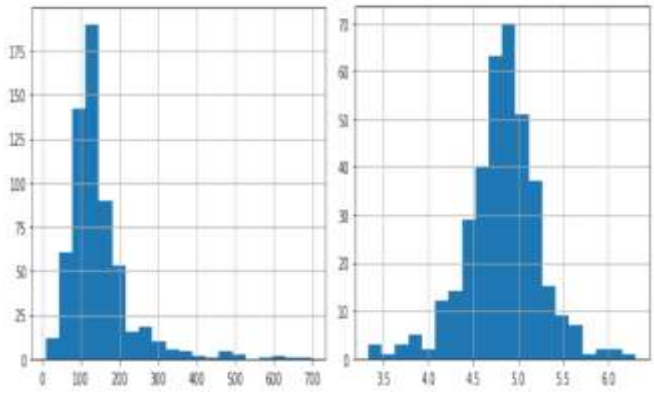

Fig 4. Log transforms

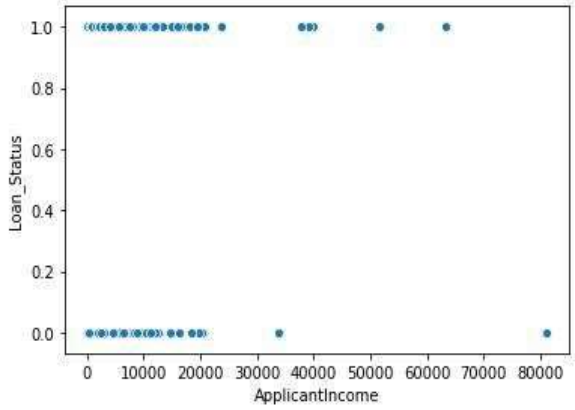

Fig. 5. Applicant ion income vs Loan Status

SVM works by mapping input data points into a higher-dimensional space using a mathematical function called a kernel. In this space, SVM tries to find the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points (support vectors) from each class. By maximizing the margin, SVM aims to achieve better generalization and robustness to unseen data.

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	614.000000	614.000000	614.000000
mean	5403.459283	1621.245798	146.412162	342.000000	0.855049
std	6109.041673	2926.248369	84.037468	64.372489	0.352339
min	150.000000	0.000000	9.000000	12.000000	0.000000
25%	2877.500000	0.000000	100.250000	360.000000	1.000000
50%	3812.500000	1188.500000	129.000000	360.000000	1.000000
75%	5795.000000	2297.250000	164.750000	360.000000	1.000000
max	81000.000000	41667.000000	700.000000	480.000000	1.000000

Table 1. distribution of numerical columns, including LoanAmount, ApplicantIncome, and others.

For linearly separable data, SVM finds the optimal hyperplane that separates the classes with the largest margin. However, in real-world scenarios where data may not be linearly separable, SVM can still perform well by using different kernel functions such as polynomial, radial basis function (RBF), or sigmoid to map the data into a higher dimensional space where separation is possible.

```
 0   Loan_ID            614 non-null    object
 1   Gender             601 non-null    object
 2   Married            611 non-null    object
 3   Dependents         599 non-null    object
 4   Education          614 non-null    object
 5   Self_Employed      582 non-null    object
 6   ApplicantIncome    614 non-null    int64
 7   CoapplicantIncome  614 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  Property_Area      614 non-null    object
 12  Loan_Status        614 non-null    object
dtypes: float64(4), int64(1), object(8)
```

Fig 6.Data Description.

Prediction of granting the loan to the customers by the bank is the proposed model. Classification is the target for developing the model and hence using Logistic Regression with sigmoid function is used for developing the model. Preprocessing is the major area of the model where it consumes more time and then Exploratory Data Analysis which is followed by Feature Engineering and then Model Selection. Feeding the two separate datasets to the model, and then preceding the model. Logistic regression is a type of statistical machine learning technique/algorithm which is used to classify the data by considering outcome variables on extreme ends and tries to make a logarithmic line that distinguishes between them. By this way prediction can be made through Logistic Regression.

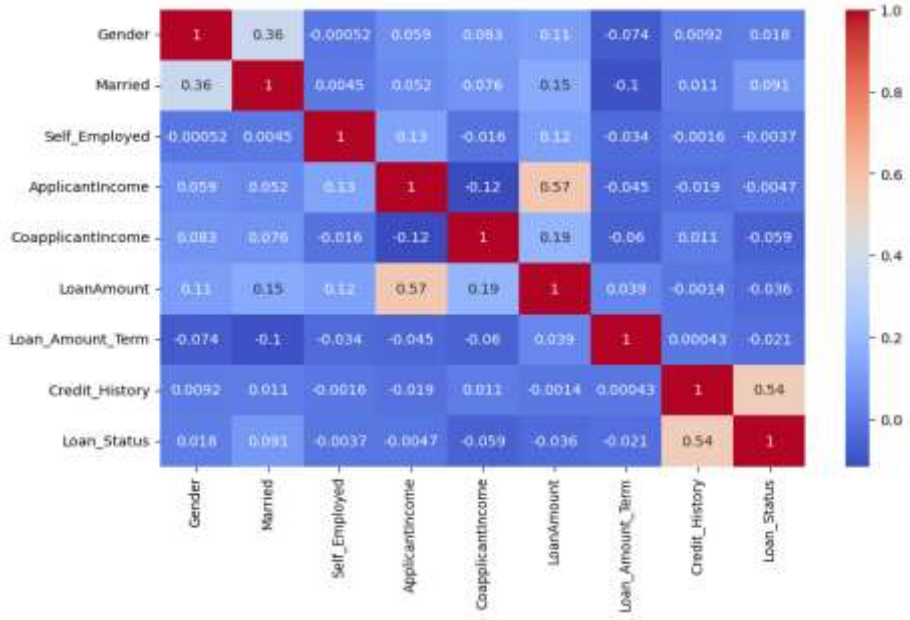

Fig. 7. Visualizing the data.

We will go each steps of the program. Firstly, Python programmers frequently use the function df.head() to show the first few rows of a DataFrame object. You can examine a preview of data in the DataFrame df by executing the function df.head(). The DataFrame df's first five rows will be printed to the console when this code is run. The head() function accepts an integer as an input if you want to display a different number of rows. For instance, df.head(10) will show the DataFrame's top ten rows.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	80.2%	79.5%	82.1%	80.8%
Decision Tree	77.5%	76.0%	78.3%	77.1%
Random Forest	84.6%	83.8%	85.7%	84.7%

XGBoost	86.1%	85.2%	87.3%	86.2%
SVM	88%	86%	88%	88%

Table 3. Comparative results.

V.CONCLUSION

In conclusion, this project successfully leverages machine learning to enhance the loan approval process by implementing and comparing the performance of Support Vector Machine (SVM) and Random Forest algorithms. By focusing on key applicant features such as credit history, income, employment status, and loan amount, the project has demonstrated the potential of these models to accurately predict loan approval outcomes. The results underscore the value of machine learning in reducing evaluation time, minimizing risks, and improving decision-making in financial institutions.

In future enhancements of the Loan Approval Prediction project, there are several avenues to explore. Integrating advanced feature engineering techniques could improve the model's predictive power by uncovering more nuanced patterns in the data. Additionally, leveraging deep learning methods, such as neural networks, could further enhance the model's ability to capture complex relationships among features. The implementation of ensemble techniques could also be considered to combine the strengths of various models, potentially leading to better overall performance.

REFERENCES

[1] Kumar Arun, Garg Ishan, Kaur Sanmeer, Loan Approval Prediction based on Machine Learning Approach.

[2] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, 'Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization': International Journal of Advanced Computer Science and Applications (Vol. 8,No. 10, 2017).

[3] Mohamed El Mohadab, Belaid Bouikhalene, Said Safi, 'Predicting rank for scientific research papers using supervised learning'Applied Computing and Informatics 15 (2019) 182–190.

[4] K. Hanumantha Rao, G. Srinivas, A. Damodhar, M. Vikas Krishna: Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: Internatinal Journal of Computer Science and Telecommunications (Volume2, Issue3, June 2011).

[5] J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086. [6] G. Arutjothi, C. Senthamarai: Prediction of loan status in commercial bank using machine learning classifier, International Conference on Intelligent Sustainable Systems (ICISS), 2017.

[7] J.R. Quinlan. Induction of decision trees. Machine learning Springer, 1(1):81–106, 1086.

[8] K I Rahmani, M.A. Ansari, Amit Kumar Goel, "An Efficient Indexing Algorithm for CBIR,"IEEE- International Conference on Computational Intelligence & Communication Technology ,13-14 Feb 2015.

[9] Gurlove Singh, Amit Kumar Goel ,"Face Detect ion and Recognition System using Digital Image Processing" , 2nd International conference on Innovative Mechanism for Industry Application ICMIA 2020, 5-7 March 2020, IEEE Publisher.

[10] Amit Kumar Goel, Kalpana Bat ra, P oonam Phogat ," Manage big data using optical networks", Journal of Stat ist ics and Management Systems " Volume 23, 2020, Issue 2, Taylors & Francis.

[11] Raj, J. S., & Ananthi, J. V., "Recurrent neural networks and nonlinear predict ion in support vector machine" Journal of Soft Comput ing Paradigm (JSCP), 1(01), 33-40, 2019.

[12] Aakanksha Saha, Tamara Denning, VivekSrikumar, Sneha Kumar Kasera. "Secrets inSource Code: Reducing False Positives using Machine Learning", 2020 International Conference on Communication Systems &Networks (COMSNETS), 2020.