LOAN PREDICTION USING MACHINE LEARNING WITH PYTHON

HEMANTH KUMAR SARISA, VARUN KHURANA, VENKAT CHANDU KOTI, NEHA GARG

[1,2,3-]STUDENT, DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
FACULTY OF ENGINEERING AND TECHNOLOGY
MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH AND STUDIES,
FARIDABAD, HARYANA, INDIA
[4]ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
FACULTY OF ENGINEERING AND TECHNOLOGY
MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH AND STUDIES,
FARIDABAD, HARYANA, INDI

Sarisahemanthkumar2508@gmail.com, varunkhurana7354@gmail.com, chandukoti668@gmail.com,
nehagarg.set@mriu.edu.in

Abstract:

Loan prediction is a significant problem in the banking industry. Using Technology, will change and make some improvements in Loan prediction by bank employees manually. By using “Machine Learning “in loan prediction. The machine learning system will check the data for a loan of the applicant and the data will be checked so fast and without any human error. So, it will help to improve technology in Bank sectors. By this, processing the data will be checked quickly, making the data output in less time. It helps Both the applicant and the bank employees.

Machine learning algorithms can be used to analyze historical data and make accurate predictions about whether a loan application will be approved or not. In this study, we propose a machine learning-based loan prediction model that utilizes various features such as credit score, income, loan amount, and loan term. We explore different classification algorithms, including logistic regression, decision tree, and random forest, to build the loan prediction model. Our experimental results show that the random forest algorithm outperforms other models with an accuracy rate of 83.45%. The proposed model can help banks and financial institutions streamline their loan approval process, reduce manual effort, and minimize the risk of bad loans.

Keywords:

Python, Machine Learning, Data Pre-Processing, Random Forest Classifier, Decision Tree, KNeighbors Classifier

1. Introduction:

Machine learning has become an increasingly popular tool for predicting loan outcomes in recent years. Lenders are always looking for ways to minimize risk and increase the chances of loan repayment. Machine learning models can be trained to analyze large datasets and predict whether or not a borrower will be able to repay a loan based on a variety of factors [1]. The goal of this research paper is to explore the use of machine learning algorithms in predicting loan outcomes. We will analyze various factors that influence loan repayment, including credit score, income level, and loan amount. We will also evaluate the performance of different machine learning models in predicting loan repayment. This research has significant practical implications for lending institutions, as it can help them make more informed decisions about which loans to approve and which borrowers to lend to. We hope that this research will contribute to a better understanding of the potential of machine learning in the lending industry.

1.1 Objective:

The main objective of this research paper on loan prediction using machine learning is to develop a predictive model [9] that can accurately predict the likelihood of loan repayment based on a variety of factors. Specifically, we aim to:

1. Identify the most important factors that influence loan repayment, such as credit score, income level, loan amount, and loan term [4].
2. Collect and analyze a large dataset of loan applications and outcomes to train and test our machine learning model.
3. Evaluate the performance of different machine learning algorithms, such as logistic regression, decision trees, and neural networks, in predicting loan repayment.
4. Compare the accuracy of our machine learning model to traditional credit scoring methods, such as FICO scores.
5. Provide recommendations for lending institutions on how to use machine learning models to improve loan approval processes and minimize risk [7].
6. Ensure that the loan prediction model complies with legal and regulatory requirements related to lending practices and anti-discrimination laws.

Overall, the objective of this research is to demonstrate the potential of machine learning in the lending industry and its ability to improve loan decision-making processes.

1.2Literaturereview:

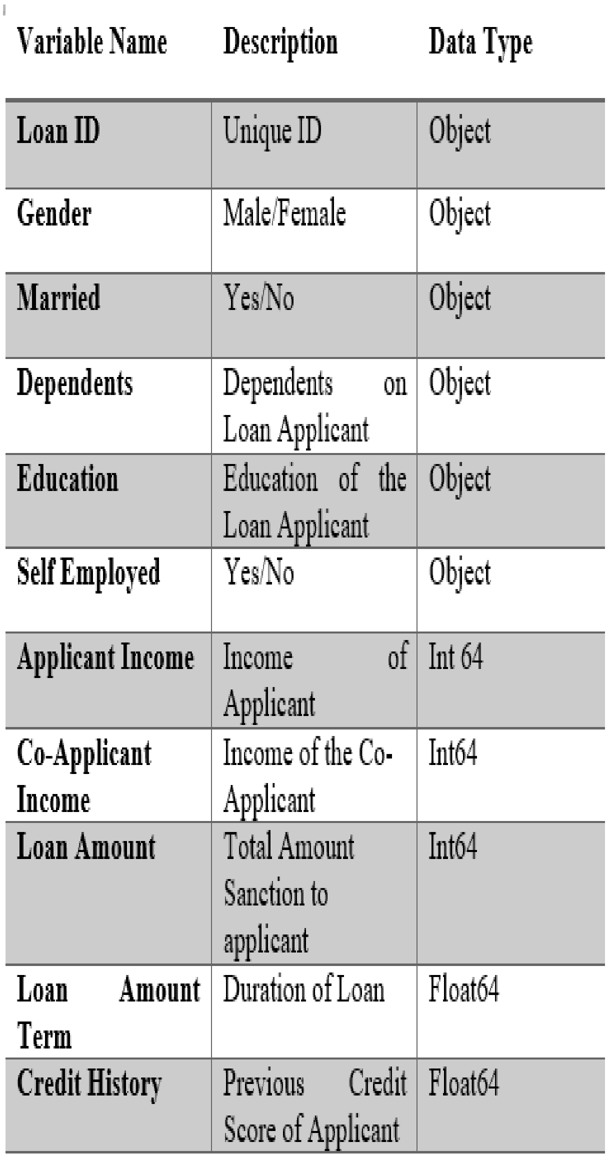
There are micro and macro factors that affect housing expenses. These components are divided into three crucial groups for this investigation: state of being, thought, action, and region [2]. The range of the house, the number of rooms, the availability of a kitchen and parking space, the openness of the yard nursery, the zone of land and structures, and the age of the house are states of being that can be observed by human

beings, while the thought is an idea offered by architects to entice potential buyers, such as the possibility of a moderate home, strong and green conditions, and world-class conditions [3]. The zone has a significant impact on how much a home costs. Loan prediction through machine learning techniques has gained significant attention in the financial industry, owing to its potential to enhance credit risk assessment [6], streamline loan approval processes, and reduce default rates. Scholars have investigated diverse components influencing loan outcomes.

The analysis categorizes these components into three essential groups: state of being, thought, and territory. State of being referred to the intrinsic properties of loan applicants, including credit history, income, employment status, debt-to-income ratio, and loan purpose. Thought encompasses factors offered by lenders to attract potential borrowers, such as attractive interest rates, flexible repayment terms, and personalized loan products [9]. Territory plays a vital role in shaping loan decisions as it determines the economic environment, market conditions, collateral, and accessibility to public amenities. Proximity to schools, hospitals, shopping centers, and recreational areas influences loan eligibility and interest rates [2].

1. Dataset:

In Loan Prediction, the Data is taken from the Data Set. The Data Set consisted of some variables which were taken from the Bank. The dataset consists of tags like Loan ID, Gender, Married, dependents, etc [2] as shown in Figure 1. The Data Set is used to take the information from the loan Applicants. With, the help of this data the machine learning algorithm will find whether the applicant applies for a loan (or) not [3].



Variable Name	Description	Data Type
Loan ID	Unique ID	Object
Gender	Male/Female	Object
Married	Yes/No	Object
Dependents	Dependents on Loan Applicant	Object
Education	Education of the Loan Applicant	Object
Self Employed	Yes/No	Object
Applicant Income	Income of Applicant	Int 64
Co-Applicant Income	Income of the Co-Applicant	Int64
Loan Amount	Total Amount Sanction to applicant	Int64
Loan Amount Term	Duration of Loan	Float64
Credit History	Previous Credit Score of Applicant	Float64

Fig: 1 Dataset Variables, their description and datatype

Machine Learning uses this data set as a training dataset. By this Dataset, the model will train with the help of this Dataset. After the training of the model, the new entries act as test data which was filled in at the time of applying. After performing the tests, the model will be able to predict whether the applicant can pay the loan (or) not.

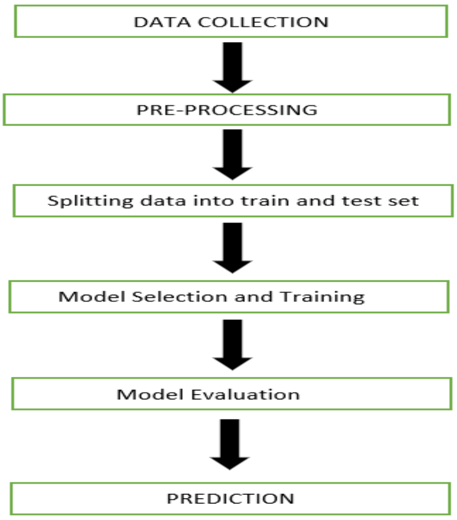

DATA COLLECTION
PRE-PROCESSING
Splitting data into train and test set
Model Selection and Training
Model Evaluation
PREDICTION

Fig:2 Steps for Loan Prediction

In the above Figure 2 the steps to predict the loan whether is approved by the loan applicant (or) not. First, we take the data set check the data, and remove the null values in the data set [11]. Then, we split the data into Train Data and test Data. Then, we apply a machine learning algorithm to find the accuracy of the data set By using different types of machine learning algorithms we get an accuracy rate. We prefer the best of all in those algorithms. By, those algorithms we can predict the loan approval.

2.1 Data collection:

Data collection refers to the process of gathering information and acquiring data from various sources. It involves systematically gathering relevant data to address specific research questions, support decision-making, or analyze patterns and trends [8]. Effective data collection requires careful planning, clear objectives, and appropriate methodologies. It also involves ensuring the accuracy, reliability, and ethical handling of data. Depending on the nature of the data, considerations must be made to protect privacy and comply with relevant data protection regulations. I have collected the various datasets from GitHub. We have taken a dataset that is suitable for loan prediction. This dataset has less scope for errors and variations [12].

2.2 Pre-Processing:

Pre-processing refers to the steps taken to clean, transform, and prepare data before analysis or modeling. It involves various techniques and operations to ensure that the data is in a suitable format for further analysis and to address issues such as missing values, outliers, noise, or inconsistencies. Pre-processing steps are highly dependent on the specific characteristics of the dataset, the analysis goals, and the modeling techniques to be applied. It is crucial to carefully assess and understand the data to determine the appropriate pre-processing steps required for a particular analysis task.

2.3 Splitting Data into Train Set & Test Set:

Splitting data into a train set and a test set is a common practice in data analysis and machine learning. The purpose of this split is to evaluate the performance of a model on unseen data and to avoid overfitting, which occurs when a model performs well on the training data but poorly on new, unseen data. Splitting the data into train and test sets helps estimate the model's ability to generalize to new, unseen data. It allows for unbiased evaluation of model performance and helps identify potential issues like overfitting or underfitting.

2.4 Model Selection & Training:

Model selection and training is a critical step in machine learning, where you choose an appropriate algorithm and train it on your dataset. It's important to choose evaluation metrics based on the problem type, data characteristics, and specific goals of your project. Consider the context, the potential impact of false positives and false negatives, and any domain-specific requirements. Additionally, keep in mind that evaluation should not be limited to a single metric, but should be a comprehensive analysis considering multiple metrics and a thorough understanding of the problem at hand.

2.5 Model Evaluation:

Model evaluation is a crucial step in machine learning to assess the performance and effectiveness of a trained model. It involves measuring how well the model generalizes to new, unseen data and how accurately it predicts the target variable. Here are some common techniques and metrics used for model evaluation. Model selection and training is an iterative process. It often involves experimentation, fine-tuning, and comparing different models to find the one that best suits your problem and data.

2.6 Prediction:

Prediction, in the context of data analysis and machine learning, refers to the process of estimating or forecasting an unknown or future outcome based on available data and a trained model. It involves using a trained model to make informed guesses or projections about what might happen in a given situation. It's important to note that the accuracy of loan predictions depends on the quality and representativeness of the training data, the chosen machine learning algorithm, and the features used in the prediction model. Regular model evaluation and monitoring are necessary to ensure its performance and to adapt to any changes in data patterns or application requirements.

3. Algorithm used in Machine Learning:

Machine Learning is a part of AI (Artificial Intelligence). By, using this machine learning we can find the accuracy of the algorithm [5] which will help to predict the loan in this project. In this project, we use three types of Machine [10] Learning algorithms which are given below:

3.1 RandomForestClassifiers

3.2 Decision Tree

3.3 KNeighborsClassifiers

3.1. Random Forest Classifiers:

RandomForestClassifier is a class in sci-kit-learn, which is a popular machine-learning library in Python. It is an implementation of the random forest algorithm for classification tasks. Random forests are an ensemble learning method that combines multiple decision trees to make predictions as shown in Figure 3. The RandomForestClassifier builds a collection of decision trees using a technique called bootstrap aggregating (or bagging) and random feature sub-selection. Each decision tree is trained on a random subset of the training data, and at each split, a random subset of features is considered. This randomness helps to reduce overfitting and improve the model's generalization capability.

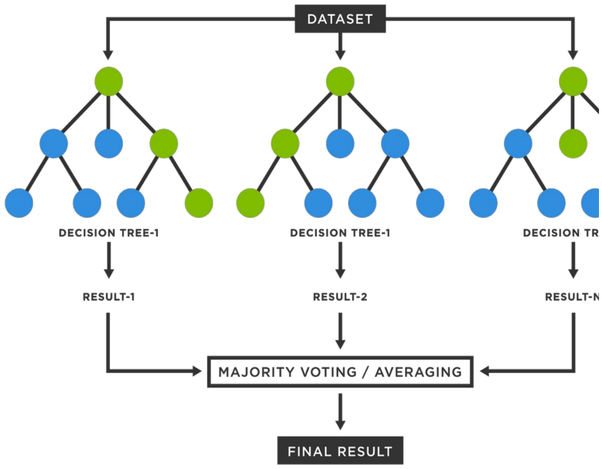


Fig: 3 Random Forest Classifier

3.2. Decision Tree:

A decision tree is another popular machine learning algorithm used for classification problems, such as loan prediction [7]. A decision tree is a tree-like model where each node represents a feature or attribute, and each edge represents a decision or rule based on that feature as shown in Figure 4. The goal of the decision tree is to split the dataset into increasingly homogeneous subsets until a stopping criterion is met [15].

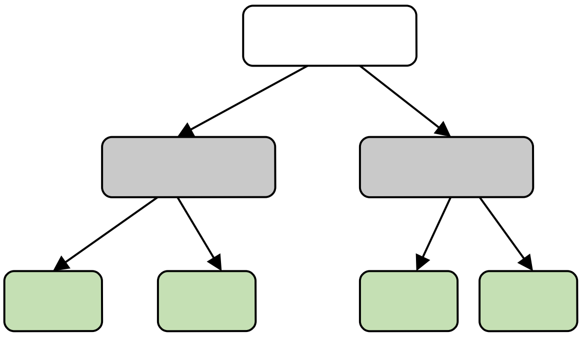


Fig:4 Decision Tree

3.3. KNeighbors Classifier:

The KNeighborsClassifier is a classification algorithm in machine learning that belongs to the category of instance-based or lazy learning algorithms. It is used for supervised learning tasks, specifically for classification problems. The KNeighborsClassifier works by classifying a new data point based on the classes of its k nearest neighbors in the feature space. As shown in Figure 5 The "k" in KNeighborsClassifier refers to the number of neighbors considered in the classification process.

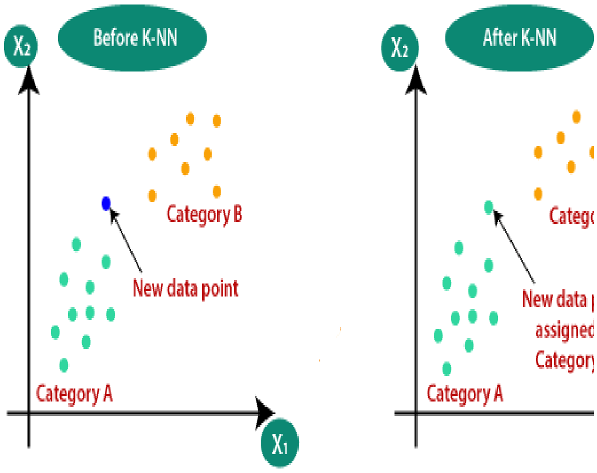


Fig: 5 KNeighbor's Classifier

4. Results:

To achieve the result for loan prediction various machine learning algorithms have been utilized. Macro and micro factors, that affect the calculation for loan prediction are considered to provide the desired result [13]. Data collecting is started first. Then, data cleaning is carried out to make the data clean and error-free. following data preparation is finished. The distribution of the data is then intended to be depicted through the creation of various graphs using data visualization. In the end, the commercial costs of the homes were calculated precisely and precisely. This was possible because our house pricing dataset's multiple regression methods were applied to improve their accuracy and produce better results. This improvement was made possible by a straightforward stacking algorithm.

In addition to applying regression techniques, some classification algorithms are also taken into account, including the decision tree algorithm, Random Forest classifier, Kneighbour classifier, etc. The Accuracy of the Machine Learning algorithm in this Loan Prediction is shown in the table below:

S.No.	Machine learning Algorithm	Accuracy
1	Random Forest Classifier	76.66%
2	Decision Tree	67.52%
3	KNeighborsClassifier	80.83%

Table 1: Statistics of ML Algorithms for Loan Prediction

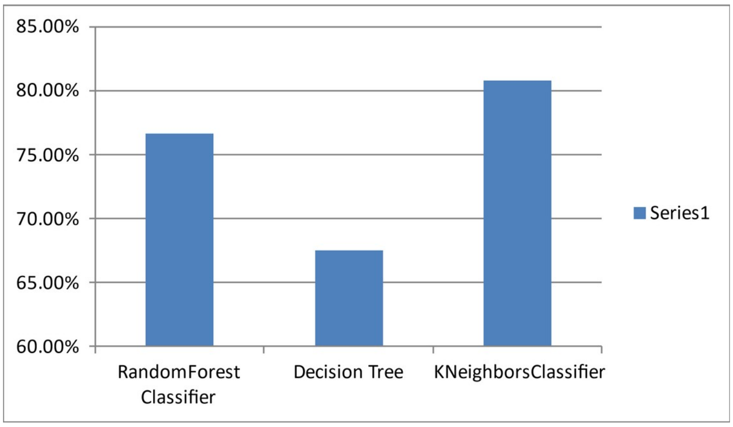


Fig:5

The above graph will represent the Percentage of the algorithms for loan prediction using machine learning we get the accuracy rate of the Random Forest Classifier is 76.66% the Decision Tree is 67.52% and the KNeighbors Classifier is 80.83%.

5. Conclusion:

For Loan prediction using a Random Forest Classifier an accuracy of 76.66% was achieved. By, using the Decision Tree Algorithm we have an accuracy of 67.52% and by using KNeighbors Classifiers we have an accuracy of 80.83%. By, considering all three algorithms the best and best is the KNeighbors Classifier because it has a good accuracy rate as compared to the remaining algorithms. KNeighborsClassifier performs the best for your specific loan prediction task, it would be the recommended choice. In conclusion, machine learning algorithms have demonstrated their effectiveness in loan prediction tasks. The comparative analysis presented in this research highlights the strengths and weaknesses of decision tree-based models, random forest classifiers, and K-nearest neighbors classifiers.

Future research should focus on exploring advanced techniques, such as deep learning models or hybrid approaches, and incorporating alternative data sources to further enhance the accuracy and robustness of loan prediction models. Additionally, investigating the ethical considerations, fairness, and transparency of machine learning algorithms in loan decision-making is essential to ensure unbiased and responsible lending practices in the financial industry.

Reference:

1. Arun, K., Ishan, G. and Sanmeet, K., 2016. Loan approval prediction based on machine learning approach. *IOSR J. Comput. Eng*, *18*(3), pp.18-21.
2. Bhattad, S., Bawane, S., Agrawal, S., Ramteke, U. and Ambhore, P.B., 2021. Loan Prediction using Machine Learning Algorithms. *International Journal of Computer Science Trends and Technology*, *9*(3), pp.143-146.Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
3. DeNicola, Louis,” What Is A Good Credit Score.” Web Blog post. Score Advise. 6 June 2023.
4. R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartotoor, and I.B. Syamwil, "Factors influencing the price of housing in Indonesia", *Int. J. Hous. Mark. Anal.*, vol. 8, no. 2, pp. 169-188, 2015.
5. P Arokianathan, V Dinesh, B Elamaran, M Veluchamy, and S Sivakumar, “Automated Toll Booth and Theft Detection System”, IEEE 2017 Technological Innovations in ICT for Agriculture and Rural Development (TIAR), pp. 84-88, 07th - 08th April 2017. DOI: 10.1109/TIAR.2017.8273691
6. Ereiz, Z., 2019, November. Predicting default loans using machine learning (OptiML). In *2019* 27th Telecommunications Forum (TELFOR) (pp. 1-4). IEEE.
7. Quinlan, J.R., 1986. Induction of decision trees. Machine learning, 1, pp.81-106.
8. Rao, K.H., Srinivas, G., Damodhar, A. and Krishna, M.V., 2011. Implementation of anomaly detection technique using machine learning algorithms. International journal of computer science and telecommunications, 2(3), pp.25-31.
9. M. Jain, H. Rajput, N. Garg and P. Chawla, "Prediction of House Pricing using Machine Learning with Python," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 570-574, doi: 10.1109/ICESC48915.2020.9155839.
10. Fan, C., Cui, Z., &Zhong, X. (2018, February). House Prices Prediction with Machine Learning Algorithms. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing (pp. 6-10).ACM.
11. Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7),1145- 1159.
12. Liu, J., Ye, Y., Shen, C., Wang, Y., &Erdélyi, R. (2018). A New Tool for CME Arrival Time Prediction using Machine Learning Algorithms: CATPUMA. The Astrophysical Journal, 855(2), 109.
13. Kadir, T., & Gleeson, F. (2018). Lung cancer prediction using machine learning

and advanced imaging techniques. Translational Lung Cancer Research, 7(3), 304-312.

14. A. Goyal and R. Kaur, “Accuracy Prediction for Loan Risk Using Machine Learning Models”.

15. J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.