Loan Eligibility Prediction using Machine Learning Algorithms

Priyanka Sananse

Assistant Professor

Excelsior K.C.College of Engineering & Management Studies and Research, Thane India

priyanka.sonawane@kccemsr.edu.in

Manasi Kesarkar

Department of Information Technology Engineering

Excelsior K.C.College of Engineering & Management Studies and Research, Thane, India

manasikesarkar@kccemsr.edu.in

Shruti More

Department of Information Technology Engineering

Excelsior K.C.College of Engineering & Management Studies and Research, Thane India

shrutimore@kccemsr.edu.in

Amrutha Nayak

Department of Information Technology Engineering

Excelsior K.C.College of Engineering & Management Studies and Research, Thane India

amruthanayak@kccemsr.edu.in

Abstract—**In today's world, due to the continuous increase in demands and needs of People which includes financial, educational growth, etc. The growth of a person requires Money which leads to rendering Loans from Financial Institutions. Loan approval is a quite hectic technique that requires lots of human effort it may also give rise to human errors. In this paper, a technology-based approach is used in terms of Machine Learning Algorithms. Loan Prediction is done using Machine learning algorithms like Support Vector Machine (SVM), Gaussian Naïve Bayes, K-Nearest Neighbor(KNN), and Decision Tree(DT). The primary goal to implement this eligibility loan prediction project is to reduce human errors and to develop system which gives significant good accuracy rate to smoothen the Loan processing stage.**

Keywords—Loan, Prediction, Machine learning, the accuracy rate

I. Introduction

Due to Digitization, every individual is reliable on to the internet for speed and quick process.In today's era, banks are boosting their financial status by providing different kinds of loans to people who are in need. Due to necessity of needs increasing day by day people tends to get Loan from these banking institution to satisfy their requirements. Logically and Practically, banks cannot provide loans to everyone as there are only limited reserves associated with each of them. So, banks must follow some standard verified process to approve the eligibility of loan, because if the one who got his/her loan approved failed to pay back his loan it may have a direct impact on the financial status of the bank. Therefore, banks provide loans only for a limited set of people who are capable of repaying their loans. But finding out who is eligible for the loan is a much more hectic and time-consuming process.

In this project, we will develop a model to predict who is eligible for a loan to reduce the risk associated with the decision process and modify the typical loan approval process into a much easier one.

II. Literature Survey

In the paper [1], researchers have use min-max normalization and Linear Regression combination to predict accuracy rate. They have been mostly focusing on quantity attributes like applicant income, credit-history, etc. In the paper [2], researchers have used KNN, Decision tree, Naïve Bayes models for prediction out of which KNN has given 0.834 accuracy which was highest. In the paper [3], researchers have given highest priority to Logistic Regression algorithm. In the papers [3][4], Applicants with Credit history not passing fails to get approved, Probably because that they have a probability of a not paying back. At-most, Applicants with high income get approved which make sense because they can easily pay back their loans. Some basic dataset attributes like gender and marital status seems not to be taken into consideration by the company. In the paper[5] ,authors have used decision tree induction algorithm to implement a model and tried to review credit scoring of mortgage loans and criteria for the applicants. Credit score helps in sanctioning of the loan. They have developed a model to predict it is safe or not for loan sanctioning and it was concluded that mostly low-income applicants receive loan approval as they are likely to repay their loan back. Dataset is collected from the Kaggle. In the paper [6], loan prediction is done using decision tree algorithm. In the paper [7], researchers have used LightGBM , XGBoost , Random forest and Logistic Regression algorithms and result produced by Random forest 0.89 accuracy was taken into consideration. In the paper [8], researchers tried to assess the credit risk for a Tunisian bank by modelling the default risk of its commercial loans. They have used a database of almost 1000 credit files between the years 2003 to 2006. The main results show variables improves the prediction quality, and the classification status is passed from 59.63 to 63.85 per cent, respectively, in the non-cash flow and cash flow models. The final output shows that the AUC criterion is 69

per cent. Using the same data, this rate is improved and which passed to 83 per cent when made use of the NN methodology. Their study is, however, incomplete in the sense that they only quantitative variables and it did not show the importance of qualitative variables based on strategic data in completing the financial status of a borrower. In the paper [9], researchers have used combination of min-max normalization and KNN algorithm is used which gives 75.08% accuracy result. R programming language is used to build KNN soring credit model. It was not clear which model is best because each model has its own specification. In the paper [10], authors have done data mining using KNN, Binning and Naïve-Bayesian classifier is used for predicting loan eligibility. A hybrid of naïve-Bayesian and K-means was used for improving the efficiency of the system.

III. PROPOSED METHODOLOGY

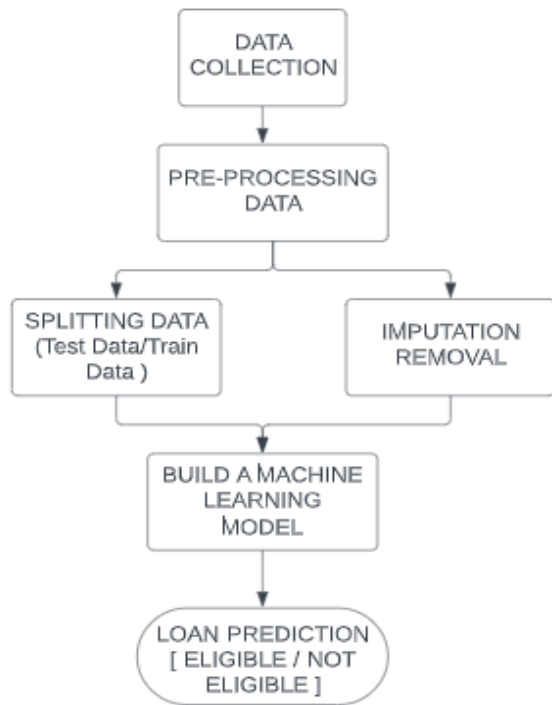The process flow of the designed loan predicting system is shown below :

Fig 1. Block Diagram

A. Process Description

There will be an application form in which applicants will need to fill in their valid credentials (Condition). Certain quantitative, categorical attributes will be taken into consideration in an application form from the dataset on which classification and clustering will be performed. Using different data analytics tools loan prediction and their severity can be cast. In this process, it is required to train the data using different algorithms and then compare user data with trained data to predict the nature of the loan. To extract some common patterns from a dataset which is chosen, and then build a model based on these extracted patterns. The training data set will be supplied to the machine learning model;

based on this data set the model is trained. Every new applicant's details filled at the time of application form acts as a test data set. After verifying valid credentials, the model predicts whether the new applicant is a fit case for approval of the loan or not. Dataset will be trained and tested using mainly four algorithms namely Naïve Bayes (Gaussian NB), Support Vector Machine (SVM), K- Nearest Neighbor (KNN), Decision Tree (DT). An algorithm that will give better accuracy and results will be selected and will be used as a Prediction Model for our system. The input to the model is the customer credentials in terms of specific attributes is collected. On the output from the classifier, a decision on whether to approve or reject the customer request can be made.

B. Machine Learning

Predictive analytics is used to predict the data about future events. Some ML type is as follows:
1) Supervised learning
2) Unsupervised Learning
In our work, we used a supervised learning approach

C. Algorithms Used

❖ Naive Bayes (Gaussian):
The classifier used: GaussianNB () with help of sklearn

❖ Support Vector Machine (SVM):
The classifier used: SVC() with help of sklearn.

❖ K -Nearest Neighbor (KNN):
The classifier used: KNeighborsClassifier() with help of sklearn.

❖ Decision Tree (DT):

Classifier used : DecisionTreeClassifier() with the help of sklearn.

IV. DATASET DESCRIPTION AND PRE-PROCESSING

We have used Kaggle Dataset which mainly consists of 13 attributes for loan Prediction as shown below :

```
> ⌄    df.info()

···   <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 614 entries, 0 to 613
      Data columns (total 13 columns):
       #   Column             Non-Null Count  Dtype
      ---  ------             --------------  -----
       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)
      memory usage: 62.5+ KB
```
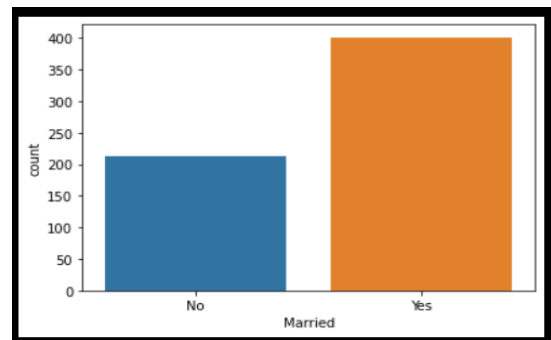
Fig 2: Attributes of datasets

We have done an Exploratory Analysis on Gender, Dependents, and Married Columns.

Exploratory Data Analysis (EDA) is an approach to analyzing the data using visual techniques. It is used to discover trends, and patterns or to check assumptions with the help of statistical summaries or graphical representations.
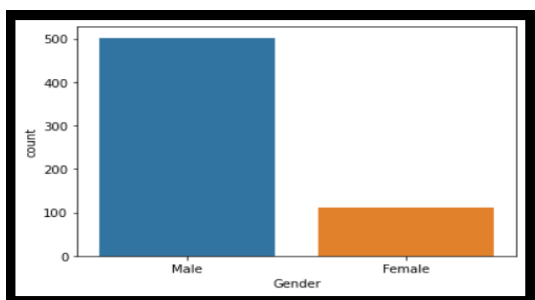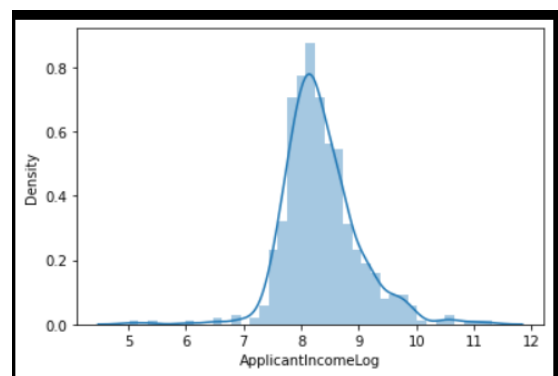
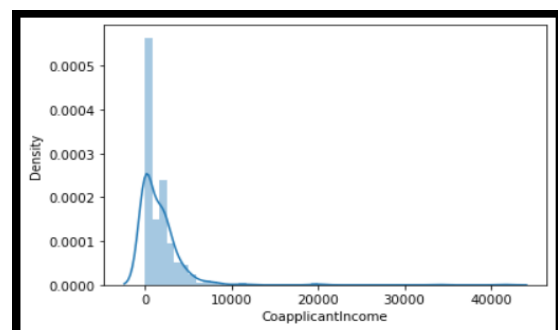*Fig 2.1 : Bar-Graph for Gender*
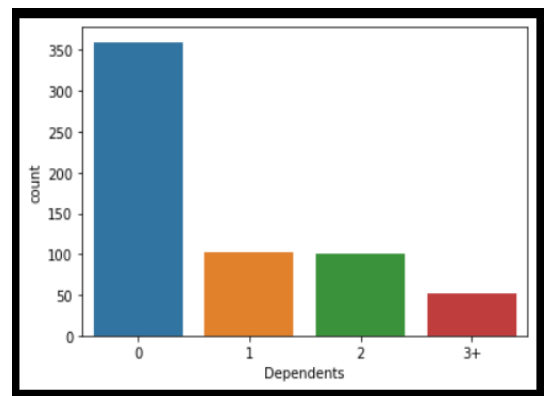
Fig 2.2 : Bar-Graph for Dependents

Fig 2.3 : Bar-Graph for Married people

For numerical data in our dataset, we have used distplot seaborn to visualize data for some attributes as shown below :

Fig 2.4 : Displot-Graph for Applicant income

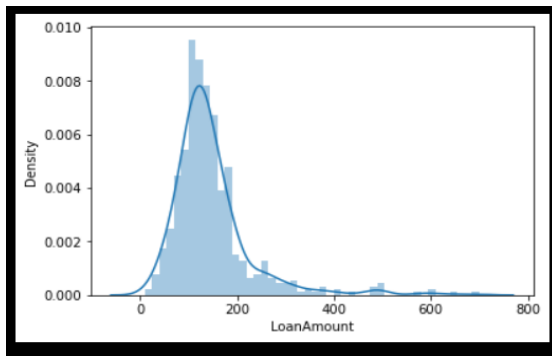*Fig 2.5 : Displot-Graph for Co-applicant income*
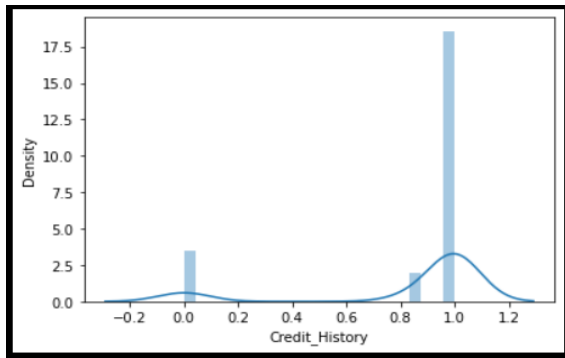
Fig 2.6 : Displot-Graph for Loan Amount

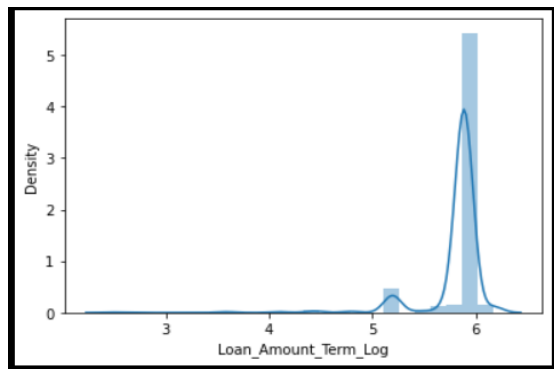*Fig 2.7 : Displot-Graph for Credit History*

Fig 2.8 : Displot-Graph for Loan amount term period

Using all the above attributes we have trained and tested data with the help of algorithms mentioned in this paper.

We have followed ML steps to get the proper model for deployment :

Firstly , we have import required machine learning libraries in order to perform execution of process smoothly .Then choose proper directory for chosen dataset in order to read data set properly. This are the task are implemented first in step known as " Directory chosen and reading train dataset".

Then, we have check in Null values in present dataset and remove them or modify them by substituting mean,median values which is termed by "Data Processing ". We have done graphical implementation on various atributes in order to analyze them easily and efficiently. We have splitted dataset into " Train" and " Test " in 80 % and 20% order. Four algorithms mentioned in papers are implemented on the dataset using "Sklearn" Library present in ML.

Finally, highest accuracy score producing algorithm is used to deploy model with the help of "Pickle" in ML In

order to extract some similar patterns ifrom dataset chosen and give reults on current prediction accurately.

V. RESULTS & CONCLUSION

After following all Machine learning steps to train and test data for proper output prediction we have used sklearn for implementing algorithms.

Below we can see the accuracy score of all the algorithms as a final result:

Fig 5 : Final results (Accuracy score rate)

Further , for project implementation final model is based on the Gaussian Naïve Bayes algorithm which is giving the highest accuracy.

We have implemented final project in form of website which is user-friendly and will quick results in instance. Below we can our system design to predict output for eligibility of users :
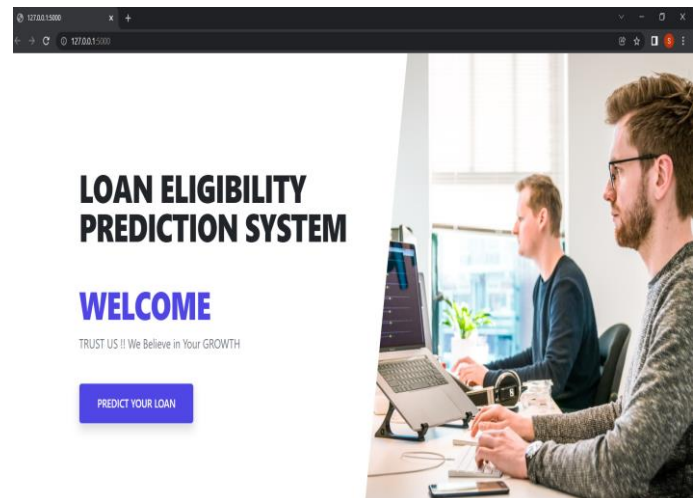

Fig 6 : Index Page of system

After clicking "PREDICT YOUR LOAN" button , user will be directed to fill a form as shown below :

Fig 7 : Application form

Here, User needs to fill all required attributes to get prediction results. Attributes to be filled are – Name of user , Gender , Marital Status , Number of dependents , Education Status , Self-employed , Credit-history , Property Area , Applicant Income , Co-applicant Income , Loan Amount and Loan Amount Term .

After filling *Fig 7 :* form , user will immediately get output of their prediction.

If user is eligible for loan then system will show result as shown below :
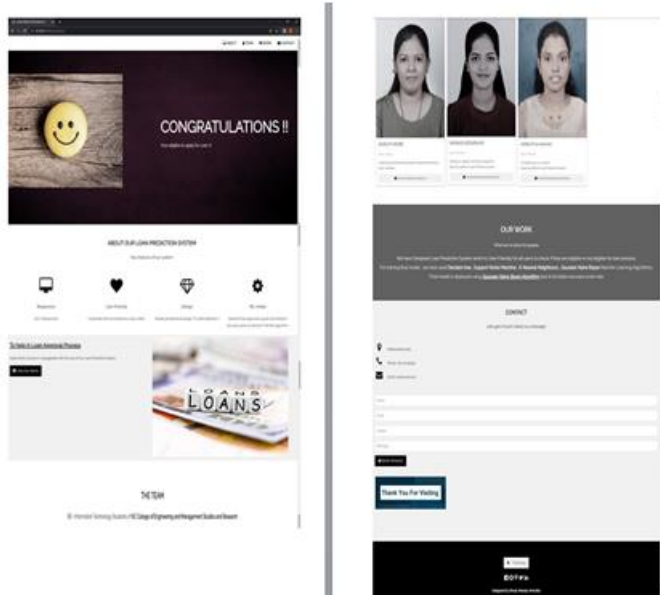

Fig 8 : Eligible Status

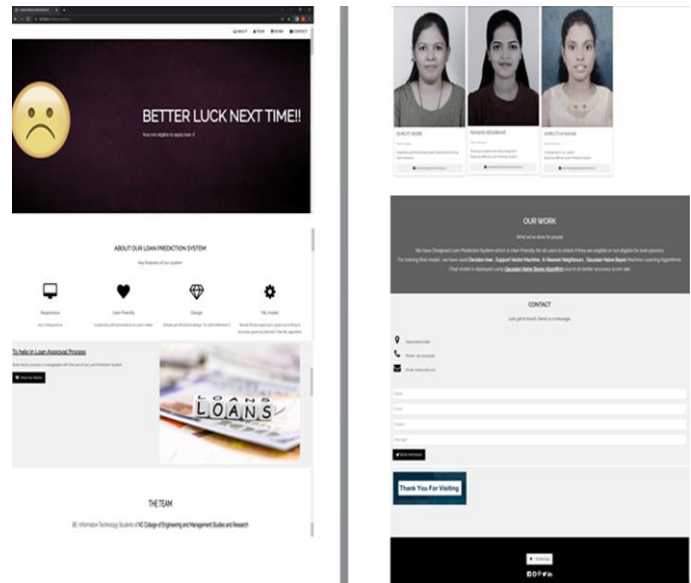If user is not eligible for loan then system will show result as shown below :

Fig 9 : Non -Eligible Status

Finally , User using this system will get their results easily and in instance of time .

FUTURE SCOPE

Predicting a person will be able to pay the debts manually is a tiring job and not always as accurate as we get when we use Machine Learning. In future, this model can be used to compare various machine learning algorithm generated prediction models .We can plan to implement activity tracking feature for project.

Using this feature the customer will be able to know the status of the loan that he/she has applied. We can integrate system to various banking institutions to smoothen their Loan approval process.Integrating this system with the bank website will add an plus point for its clients in order to check their Eligibility for loan.

This system can be used in order to differentiate loans into various categories like Housing Loan, Car Loan, Home Loan , Education Loan, etc.

REFERENCES

[1]   S. Sreesouthry, A. Ayubkhan, M. Mohamed Rizwan, D. Lokesh, K. Prithivi Raj. (2021). Loan Prediction Using Logistic Regression in Machine Learning.

[2]   Veeraballi Nagajyothi,t opic- "Loan approval prediction using KNN, decision Tree and Naïve Bayes models", International Journal of Engineering in Computer Science 2020 .

[3]   M. A. Sheikh, A. K. Goel, and T. Kumar, topic- "An Approach for Prediction of Loan Approval using Machine Learning Algorithms," At International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.

[4] PidikitiSupriya, MyneediPavani, NagarapuSaisushma, NamburiVimalaKumari, k Vikash, "Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques.Volume 5 Issue 2, Mar-Apr2019.

[5] Nikhil Madane, SiddharthNanda, "Loan Prediction using Naïve Bayes", Journal of the Gujrat Research History, Volume 21 Issue 14s, December 2019.

[6] OmPrakashYadav, ChandanSoni, SaikumarKandakatla, ShantanuSswanth, "Loan Prediction using Machine learning", International Journal of Information and Computer Science, Volume 6, Issue 5, May 2019.

[7] Coser, alexandru&maer-matei, monica&albu, crisan. (2019). predictive models for loan default risk assessment ,economic computation and Economics Cybernetics studies and research. 53. 149-165.10.24818/18423264/53.2.19.09.

[8] Abdelmoula, Aida. (2017). Using a naive Bayesian classifier methods for loan risk assessments : Evidence from a Tunisian commercial bank. Journal of Economics, Finance and Administrative Science. 22. 3-24.10.1108/JEFAS-02-2017-0039.

[9] G. Arutjothi, Dr. C. Senthamarai , topic - "Prediction of Loan Status in Commercial Bank using Machine Learning , "-2017.

[10] Aditikacheria, NidhiShivakumar, ShreyaSawker, Archana Gupta, "Loan sanctioning prediction system", International Journal of soft computing and engineering(IJSCE)- Volume-6, issue-4, September 2016.

[11] Andres C. Muller , Sarah Guido – Introduction to Machine learning using Python , First Edition , O Reilly Media , Inc. Publication.

[12] Oliver Theobald – Machine Learning for Absolute Beginners , Third Edition , Independently Published .