Loan Eligibility Prediction using Machine Learning Techniques

Mohammad Tariqul Islam Tuhin
Dept. of Computer Science and Engineering
BRAC University, Dhaka,Bangladesh
tariqul.islam.tuhin@g.bracu.ac.bd

Al Hasib Mahamud
Dept. of Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
al.hasib.mahamud@g.bracu.ac.bd

Arnab Mitra Utsab
Dept. of Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
arnob.mitra.utsab@g.bracu.ac.bd

Marwa Nabi
Dept. of Computer Scienceamd Engineering
BRAC University, Dhaka,Bangladesh
marwa.nabi@g.bracu.ac.bd

Mohammed Julfikar Ali Mahbub
Dept. of Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
mohammed.julfikar.ali.mahbub@g.bracu.ac.bd

Md Motahar Mahtab
Dept. of Computer Science and Engineering
BRAC University, Dhaka, Bangladesh
md.motahar.mahtab@g.bracu.ac.bd

Annajiat Alim Rasel
Dept. of Computer Science and Engineering
BRAC University,Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*— It can be seen that the value of assets is rising on a daily basis. That often necessary a lot of capital to buy whole asset. Many times it becomes impossible to buy even from our savings. So we can apply for a loan to get the money we need because through the loan application we can easily get the money for buying assets. Getting a loan is considered as a lengthy procedure. Before being accepted for a loan, the application have to go through a number of steps, yet approval is not sure. Many loan prediction models have been created to reduce the time and also reduce the risk attached with the loan. The main goal of our project was to compare different type of prediction model and then determine which one is the best for loan eligible prediction with lowest amount of mistake.**

Keywords*— *Machine Learning, Loan Eligibility Prediction.

I. INTRODUCTION

Prediction of loan eligibility is considered beneficial to both bank employees and applicants. The purpose of this project is to propose an efficient method to select applicants as eligible from a number of applicants. At present loan is authorized manually by a bank representative. As a result this bank employee are responsible for all things associated with it. such as he/she is eligible or not for the loan . The bank employee also determine if here have any risk or not. It takes a long time and is prone to mistakes because it is handled by a person. If the loan is not returned, the bank suffers a loss, and the interest paid to them accounts for the majority of the bank's profits. There will be a banking crisis if the banks lose too much money. The financial crisis has a negative impact on the country's economy. As a result, it is essential for the loan is authorized with the lowest number of mistake in risk prediction and in the shortest period of time feasible. As a result, a loan prediction model is needed. Because these type of model can predict that an applicant loan will be approved or not promptly. These type of model predict with the least error[1]. The project's aim is to evaluate loan eligibility prediction using various machine learning algorithms and then select the best one that may reduce loan approval time and also risk. It has been completed by forecasting whether or not the loan can be provided to that individual based on a variety of factors such as Married, Education, Loan Amount, Applicant Income, and so on. The prediction model benefits both the applicant as well as the bank through reducing risk and lowering defaulters number.

II. PROBLEM STATEMENT

Banks and other lending institutions need to know that they will get their money back, plus interest, when they approve monetary loans. As a result, before lending money, they must determine the borrower's credibility. To do so, lending authorities must extensively investigate the borrower's past and credibility. Manually going through many variables and aspects for each

loan, on the other hand, is a time-consuming and inefficient procedure.

III. LITERATURE REVIEW

In their work, Regina Esi Turkson et al. [2] look at the various machine learning models that may be used to forecast a borrower's loan eligibility. They ran the data through 15 different learning algorithms to see which ones are better for analyzing bank credit data sets. Some of the methods utilized different machine learning algorithms, the rest of the algorithms perform admirably in terms of accuracy and other performance assessment criteria, according to the experiment.

Ashlesha Vaidy [3] discusses how the technological world is fast moving toward total automation, the importance of automation, and the role of AI and Machine Learning in this process. The capacity of machines to make decisions is one of the most essential aspects to consider throughout this move to automation. Predictive and probabilistic techniques can be used to make decisions, according to the author.

Different Machine Learning algorithms are used to create them. To clarify further, the study recommends utilizing Logistic Regression to implement this predictive and stochastic method. Mohammad Ahmad Sheikh et al. [4] proved that Logistic Regression Model is a better approach to determine a borrower's creditworthiness.

The authors utilize the Logistic Regression Model to train and evaluate various factors, including the business valuation, the applicant's assets, the applicant's individual income, and so on. They contrasted these results to those produced by bank systems, which are primarily concerned with account information (which shows the wealth of the applicant). After conducting tests and comparing results, the authors discovered that the Logistic Regression Model generated marginally superior results owing to the inclusion of numerous additional characteristics such as age, purpose, credit history, credit term, and so on.

IV. PROPOSED SYSTEM

The proposed system automating the process of evaluating the creditworthiness of a potential applicant.

The details of the loan applicants are gathered into a data collection. It is organized and analyzed with the use of appropriate analytical tools. There are two types of data in this set:

- Train data is utilized to train the model, which means our model will learn from it.
- The train data contains all of the independent factors and also have the label or target variable.
- The test data does not contain the target variable. For the test data, we use the model to predict target variable.

To predict the accuracy, some machine learning method might be employed. Logistic Regression, KNN, SVM, Decision Tree, and Random Forest. And this system predicts whether or not he/she is eligible. They predict a binary result. Then, based on their accuracy, we compare these models to see which one is best for prediction. Here also XAI method use on this machine learning algorithm for explain accurately he/she why is not eligible or why he/she is eligible based on feature.

V. SYSTEM DESIGN AND IMPLEMENTATION

A. Data Description

The dataset is collect from the kaggle. In our datasets, there are 13 Columns and 614 rows available. Table 1 describes the dataset key information.

Table.1. List of attributes

Attributes	Description
Gender	Male/Female
Married	Yes/No
Self_Employed	Yes/No
Credit_History	Meets Guidelines or Not
Loan_Status	Loan Approved or Not
Dependents	Number of Dependents
Education	Graduate/Under-Graduate
Property_Area	Urban/Semi-Urban/Rural
ApplicantIncome	Applicant's Income
CoapplicantIncome	Co-Applicant's Income
LoanAmount	Loan Amount in Thousands
Loan_Amount_Term	Term of Loan in Months

Fig. 1 show a sample of the dataset. The statistical analysis of the variables is performed, and a sample is show in Fig. 2.

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y

Figure.1. Data Sample

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

Figure.2. Statistical Analysis of Data

B. Data analysis

In the dataset we can see that there are some object type of data some int type of data and some float type of data are existing in the dataset. In the figure.3 we see that how many row and column are existing in the dataset and also what type of data in our dataset.

```
<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
```

Figure.3.Data info

In our dataset there are some missing value. Now in the figure.4 we can see that some of the column have some missing value.

```
Loan_ID              0
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
Loan_Amount_Term    14
Credit_History      50
Property_Area        0
Loan_Status          0
dtype: int64
```

Figure.4.Missing value in the dataset.

C. Data Processing

The results of the distributional analysis given above are as follows:

• The data set contains some missing values.so missing value must be filled into the values.

Before the data can be fitted into the model, some extreme values or outliers must be addressed.

This problems can be handled by the following methods:

Handling missing values – For gender and married feature we drop the row where he find the missing value. And also other column such as Dependents and Credit History here also some null value. Which replace by zero. And in LoanAmount replace the missing value using median. And at last LoanAmount_Term here also some missing value. So here need to using mode to fill the missing value.

Then we need to convert categorical value into numerical value. So using level encoder we convert the categorical value into numerical value. Also there are also min max scalar method which mainly normalize the value into 0-1.It is the most basic approach. In figure.5 show the data sample after processing.

After convert the categorical value and normalize the value, then we need to define target variable which is loan_status and also from train data we drop some feature. which is not require for prediction. After that we need to split the dataset. So using train_test_split to split the dataset. So in our dataset we use train_size=0.75. That's mean 75% of the data will be used for training the model and 25% data will be used for testing the model.

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	0	1	0	0.000000	0	0	0.070489	0.000000	0.184087	0.74359	1.0	2	1
1	1	1	1	0.333333	0	0	0.054830	0.036192	0.166647	0.74359	1.0	0	0
2	2	1	1	0.000000	0	1	0.035250	0.000000	0.080924	0.74359	1.0	2	1
3	3	1	1	0.000000	1	0	0.030093	0.056592	0.173167	0.74359	1.0	2	1
4	4	1	0	0.000000	0	0	0.072356	0.000000	0.205928	0.74359	1.0	2	1

Figure.5. Data sample after processing

D. System Design

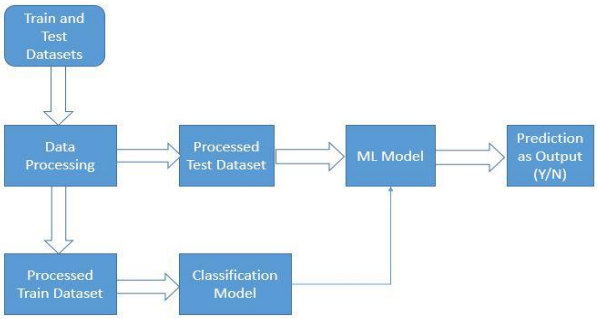

Figure.6.System Architecture

• All raw data is sent into the system, which cleans and processes it in preparation for exact testing and training.
• The classification model is fed the train data set, which defines the model's accuracy.
• The machine learning model is fed the test data set.
• Lastly, the Machine Learning Model forecasts that whether applicant is deserving of a monetary loan depending on the data provided by the classification model and the train dataset.

E. Machine Learning Model

a. Logistic Regression Model

The Python Scikit-learn package provides the Logistic Regression model. The logistic model has been used in statistics to describe the likelihood of a specific class or outcome, like pass/fail, healthy/sick, alive/dead, and win/lose.
Logistic regression is also a statistical model that utilizes a logistic function to describe a binary dependent variable in its simplest form, however there are much more complicated variants. Logistic regression is a method of calculating the parameters of a logistic model. It calculating the parameters in regression analysis. In our project, only the features that have direct impact on the loan eligibility of the applicant are considered. And other feature are drop such as Loan_id, Gender, and Loan Status. After that fit into the train model, then predict the score. In test data we drop Loan id and gender feature for predict the test data by logistic regression.
The model is made to work as follows:
•AT first train data set is first fit into the model
•Then based on the provided attributes the outputs are predicted
•The accuracy is then calculated by comparing the outputs to the train data set's previously existing outcomes.
• If the accuracy is satisfactory, the test data set is subsequently fitted into the model to produce correct predictions for each application.

b. K Nearest Neighbor (KNN)

K Nearest Neighbor (KNN) is generally a machine learning method. It is a very easy method. And it is also easy in comprehend, and adaptable. It is applied in many applications. Handwritten analysis, financial, picture classification and medical are some of the uses for KNN. Banking firms can determine a client's credit rating. During distribution of loan, banking organization predict that the loan is secure or not secure.
K Nearest Neighbor is a non-parametric learning method "Non-parametric" refers that allocation of inherent data there have no implications. Another way we can said that, from dataset, the model framework is derived. So at first train data set is first fit into the model. Then based on the provided attributes the outputs are predicted. The accuracy is then calculated by comparing the outputs to the train data set's previously existing outcomes. If the accuracy is satisfactory, the test data set is subsequently fitted into the model to produce correct predictions.

c. Support Vector Machine(SVM)

Support Vector Machine is a classification method. The primary goal in SVM is to separate the provided dataset by picking a hyperplane. Generally, it is picked in the provided dataset with the largest feasible margin among support vectors. In this project, the considered features are only those which have direct impact on the loan eligibility prediction. And other feature are drop such as Loan_id, Gender, and Loan_Status. After that fit into the train model. Then predict the score .in test data, we drop Loan id and gender features for predict the test data by SVM.

d. Decision Tree

In Decision Tree model decision rule is indicated through the branch. The outcome is indicated through every leaf node. Here the root node is the topmost node. Basically it learns to divide depending on feature values. In our project some feature are drop such as Loan_id, Gender, and Loan_Status as these features do not have any direct impact. After that fit into the train model. Then predict the score from train data. In test data, we drop Loan id and gender features for predict the test data.

e. Random Forest

Random forest is a supervised learning method which can be applied for classification and also regression. Random forest is the most adaptable and also user-friendly algorithm. How strong the forest will be depends on how many trees there are.

VI.RESULT

When using several classifiers, we can observe that Logistic Regression and SVM provide the best accuracy than other machine learning algorithms. The accuracy achieved by using several machine learning models is shown in the Table.2. Figure 12, which shows the accuracy of several classifiers in a bar chart. As a result of this analysis, it is obvious that Logistic Regression and SVM perform well in predicting loan eligibility. And also we use XAI method for explanation why the applicant are eligible and also not eligible for the loan.

Table2.Accuracy

ML algorithm	Accuracy
Logistic Regression	0.7866666666666666
K Neighbors Classifier	0.7066666666666667
SVM	0.7866666666666666
Decision Tree	0.6666666666666666
Random Forest	0.7466666666666667

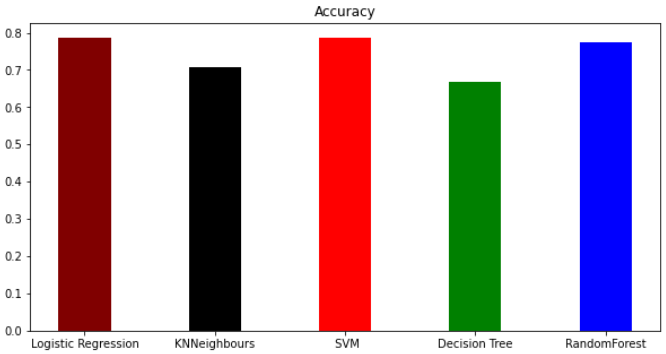

Figure.7. Bar plot of Accuracy

VII. CONCLUSION

As a consequence, it can be confidently concluded that the Logistic Regression and SVM models are very efficient and produces superior results than other models. It functions properly and satisfies all banking criteria. This system calculates the outcome correctly and precisely. It accurately predicts whether a loan will be approved or rejected for a loan application or client. Also, lime explain what is the reason for loan eligible or not.so we can easily understand what is the reason behind prediction.

VIII.REFERENCES

[1] A. Khan, E. Bhadola, A. Kumar, and N. Singh, "Loan Approval Prediction Model a Comparative Analysis," vol. 20, no. 3, pp. 427–435, 2021.

[2] O. A. Egaji, S. Ballard-Smith, I. Asghar, and M. Griffiths, "A Machine Learning Approach for Predicting Bank Credit Worthiness," *ACM Int. Conf. Proceeding Ser.*, pp. 20–24, 2020, doi: 10.1145/3418981.3418983.

[3] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," *8th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2017*, 2017, doi: 10.1109/ICCCNT.2017.8203946.

[4] M. A. Sheikh, A. K. Goel, and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," *Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020*, no. Icesc, pp. 490–494, 2020, doi: 10.1109/ICESC48915.2020.9155614.