ZKG INTERNATIONAL

ISSN: 2366-1313

# LOAN APPROVAL PREDICTION USING DECISION TREE

**[1]Mrs. T. MADHUMATHI, [2]THEEPIREDDY HAMPI, [3]SHAIK ADNAN HUSSAIN, [4]BEECHU SRIDHAR REDDY**

[1]Assistant Professor, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad, madhuthakur.26@gmail.com

[2]BTech student, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad,

hampitheepireddy@gmail.com

[3]BTech student, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad,

adnanshaik2026@gmail.com

[4]BTech student, Dept.of IT, TKR College of Engineering & Technology, Meerpet, Hyderabad,

sreddybeechu@gmail.com

***Abstract***: *Banks are making major part of profits through loans. Though lot of people are applying for loans. It's hard to select the genuine applicant, who will repay the loan. While doing the process manually, lot of misconception may happen to select the genuine applicant. Therefore, we are developing loan prediction system using machine learning, so the system automatically selects the eligible candidates. This is helpful to both bank staff and applicant. The time period for the sanction of loan will be drastically reduced. In this project we are predicting the loan data by using some machine learning algorithms that is Decision Tree.*

***Keywords***: *Machine learning, Decision tree, Banking, Load prediction, classification.*

## I. INTRODUCTION

The growing volumes, varieties and velocity of data due to the emergence of the Internet in particular and the cheaper data sharing and storage facilities coupled with the cheaper but more powerful computational tools have opened a new frontier in the field of data science. And thus, there is currently active ongoing research within the fields of data mining (discovering patterns in data) and machine learning's (building analytical models using algorithms for machine to "learn" from data), both aim at using algorithms and concepts to extract knowledge and pattern from data [1].

One of the major reasons for establishing banks is to advance loans to customers. But in order to stay in business, banks advance these loans to people who have the ability to pay back the money, thereby minimizing the risk of the non- payments of loans. However, risk management; knowing who is credit worthy is still an on-going challenge within the banking sector. The ability to identify a risk score of a customer base on some features such as occupation, age, marital status, salary range/amount of equity, credit history, etc. is an important step that banks go through before giving credit to customers. For the credit risk score helps the banks to decide on how much interest to charge on the loan, etc. However, these risk factors sometimes do not give an inform decision on the credit worthiness of customers. Moreover, many banks lack a central well integrated, automated finance and risk management system due to the inability to develop a robust and scalable risk management system to forecast risk score of customers [2].

Another nightmare faced by many banks these days is frauds. And the machine learning approach is seen and considered as the right tool that can be leveraged on in order to understand the banking transaction pattern of customers, by identifying pattern in customer data, so as to be able to distinguish between fraudulent activity from that of a normal one . Therefore, we leveraged it on the bank credit dataset in order to understand the key factors that influence the payment of bank loans. The dataset is obtained from the UCL machine repository. We perform analysis and applied machine learning algorithms on the bank credit data, firstly to understand the nature of the data and the best algorithms suitable for learning bank credit data. Secondly, we determined among the 23 features of bank customers, which ones are the most important in determining the credit worthiness of a customer. Thirdly, we formulated a predictive model to determine the credit worthiness or otherwise of a given bank customer using a linear regression method [3].

Nowadays, Banks are struggling a lot to get an upper edge over each other to enhance overall business due to tight competition. Most of the banks have now realized that retaining the customers and preventing fraud must be the strategy tool for a healthy competition. Availability of the huge quantity of data, creation of knowledge base and efficient utilization of the same have helped banks to open up efficient delivery channels. Business decisions can be optimized very well through data mining. Credit scoring, Customer segmentation, predicting payment from customers, marketing, detecting fraud transactions, cash management and forecasting operations, optimizing stock portfolios and ranking investments are some of the areas where data mining techniques can be very useful and can be used widely in the banking industry. Credit risks which result for the risk of loss and loan defaults are the major source of risk encountered by banking industry.

Data mining techniques such as classification and prediction can be applied to overcome this to a great extent. There are mainly two objectives that can be achieved through these techniques. They are as follows: 1) Identification of the relevant attributes that indicate the capacity of the borrowers to pay back the loan, and 2) Determining the best model to evaluate loan risk. Decision Tree Algorithm is one of the best techniques to achieve this objective. The model thus developed will provide a better loan risk assessment, which will potentially lead to a better allocation of the bank’s capital. In this regard, a study is conducted and an efficient prediction model which helps to reduce the proportion of unsafe borrowers is introduced herewith. Due to the significance of loan risk analysis, this study helps banking industry by providing additional information to the loan decision-making process, potentially decreases the cost and time of loan applications approving process, and decreases the level of uncertainty for loan officers by providing them enough knowledge extracted from previous loans. Decision Tree Algorithm used in this model is the data mining technique for predicting credibility of customers [4].

## II. LITERATURE SURVEY

Data mining techniques are really becoming very popular nowadays because of the wide availability of large quantity of data and the need for transforming such data into knowledge. Techniques of data mining are implemented in various domains such as retail industry, telecommunication industry, intrusion detection and other scientific applications. Data mining techniques can also be used in the banking industry which help them compete in the market well equipped. In this paper we have introduced an effective prediction model for the bankers that help them predict the credible customers who have applied for loan. Decision Tree Data Mining Algorithm is applied to predict the attributes relevant for credibility. A prototype of the model has been described in this paper which can be used by the organizations in making the right decision to approve or reject the loan request of the customers.

In[5] Banking Industry always needs a more accurate predictive modelling system for many issues. Predicting credit defaulters is a difficult task for the banking industry. The loan status is one of the quality indicators of the loan. It doesn't show everything immediately, but it is a first step of the loan lending process. The loan status is used for creating a credit scoring model. The credit scoring model is used for accurate analysis of credit data to find defaulters and valid customers. The objective of this paper is to create a credit scoring model for credit data. Various machine learning techniques are used to develop the financial credit scoring model. In this project, we propose a machine learning classifier-based analysis model for credit data. We use the combination of Min-Max normalization and K Nearest Neighbor (K-NN) classifier. The objective is implemented using the software package R tool. This proposed model provides the important information with the highest accuracy. It is used to predict the loan status in commercial banks using machine learning classifier.

In [6] National student loans have the general features of commercial loans, and are a financial credit services provided by commercial banks. But the general personal credit rating assessment system of commercial bank cannot make the correct credit rating because the lender, college students, have no credit history. To avoid the credit risk, a rational credit assessment system must to be established for college Students. With the self-learning, self-organizing, adaptive and nonlinear dynamic handling characteristics of Artificial Neural Network, a Back Propagation neural network was developed to evaluate the credit rating about a college student. Several samples, which were provided by a bank, were used for network training and testing by MATLAB. The maximum value of the error between the prediction value of the network and actual value is only 2.92 that the algorithm developed is fairly efficient for the assessment about the college student's personal credit situation.

In [7] Fraudulent activities in financial institutes can break the economic system of the country. These activities can be identified using clustering and classification algorithms. Effectiveness of these algorithms depend on quality of the input data. Moreover, financial data comes from various sources and forms such as financial statements, stakeholders' activities and others. This data from various sources is very vast and unstructured big data. Hence, parallel distributed pre-processing is very significant to improve the quality of the data. Objective of this work is dimensional reduction considering feature selection and extraction algorithm for large volume of financial data. In this paper an attempt is made to understand the implications of feature extraction and transformation algorithm using Principal Feature Analysis on the financial data. Effect of reduced dimension is studied on various classification algorithms for financial loan data. Parallel and distributed implementation is carried out on IBM Bluemix cloud platform with spark notebook. The results show that reduction of features has significantly improved execution time without compromising the accuracy.

In[8] Networked-guarantee loans may cause the systemic risk related concern of the government and banks in China. The prediction of default of enterprise loans is a typical extremely imbalanced prediction problem, and the networked-guarantee make this problem more difficult to solve. Since the guaranteed loan is a debt obligation promise, if one enterprise in the guarantee network falls into a financial crisis, the debt risk may spread like a virus across the guarantee network, even lead to a systemic financial crisis. In this pro, we propose an imbalanced network risk diffusion model to forecast the enterprise default risk in a short future. Positive weighted k-nearest neighbours (pwkNN) algorithm is developed for the stand-alone case – when there is no default contagious; then a data-driven default diffusion model is integrated to further improve the prediction accuracy. We perform the empirical study on a real-world three years loan record from a major commercial bank. The results show that our proposed method outperforms conventional credit risk methods in terms of AUC. In summary, our quantitative risk evaluation model shows promising prediction performance on real-world data, which could be useful to both regulators and stakeholders.

## III. EXISTING SYSTEM

Bank employees check the details of applicant manually and give the loan to eligible applicant. Checking the details of all applicants takes lot of time. Banks need to analyse for the person who applies for the loan will repay the loan or not. Sometime it happens that customer has provided partial data to the bank, in this case person may get the loan without proper verification and bank may end up with loss. Bankers cannot analyse the huge amounts of data manually; it may become a big headache to check whether a person will repay its loan or not. It is very much necessary to know the person getting loan is going in safe hand or not. So, it is pretty much important to have an automated model which should predict the customer getting the loan will repay the loan or not.

### Drawbacks

- Checking details of all applicants consumes lot of time and efforts.
- There is chances of human error may occur due checking all details manually.
- There is possibility of assigning loan to ineligible applicant

## IV. PROPOSED SYSTEM

To deal with the problem, we developed automatic loan prediction using machine learning techniques. We will train the machine with previous dataset. so, machine can analyse and understand the process. Then machine will check for eligible applicant and give us result.

## ALGORITHM USED

### Decision Tree

Decision Tree is a supervised learning algorithm used to solve classification and regression problems too. Here, DT uses tree representation to solve the prediction problem, i.e., external node and leaf node in a tree represents attribute and class labels respectively. Bank customers between the ages of 20 and 60 years with small limited bank balance are seen to be the highest defaulters in paying their bank loans.

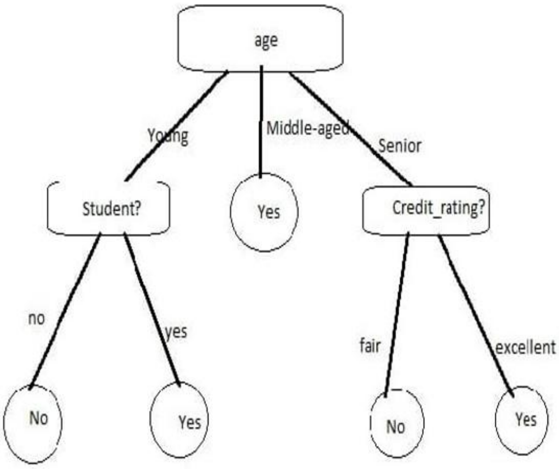


Fig.1 Decision tree algorithm

## MODULE DESCRIPTION

### Admin

Admin can login to the Application and add to the credit holder Information and his limit to use. And admin can monitor the suspects and he can block/activate users.

### User

User can register to the application and he can perform the activities of credit card usage like buy products, pay money.

### Data Collection

The dataset collected for predicting loan default customers is predicted into Training set and testing set. Generally, 80:20 ratios are applied to split the training set and testing set. The data model which was created using Decision tree is applied on the training set and based on the test result accuracy, Test set prediction is done. Following are the attributes

### Pre processing

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done. In order to overcoming these issues we use map function.

### Correlating attributes

Based on the correlation among attributes it was observed more likely to pay back their loans. The attributes that are individual and significant can include Property area, education, loan amount, and lastly credit History, which is since by intuition it is considered as important. The correlation among attributes can be identified using corplot and box plot in Python platform.

### Building the classification

Model using Classification Algorithms for predicting the loan defaulter's and non defaulter's problem LSTM algorithm is used. It is effective because it provides better results in classification problem. It is extremely intuitive, easy to implement and provide interpretable predictions.

It produces out of bag estimated error which was proven to be unbiased in many tests. It is relatively easy to tune with. It gives highest accuracy result for the problem.

### Predicting default outcomes

We noticed that 299 cases in the test set are predicted as "Y", which is more than 81%, whereas in the training set only about 69% had this status.

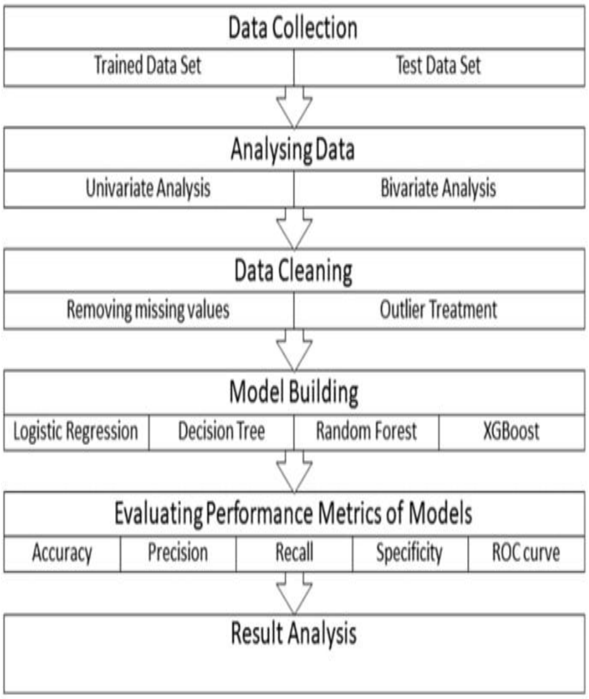


Fig.2 Module process

## SYSTEM ARCHITECTURE

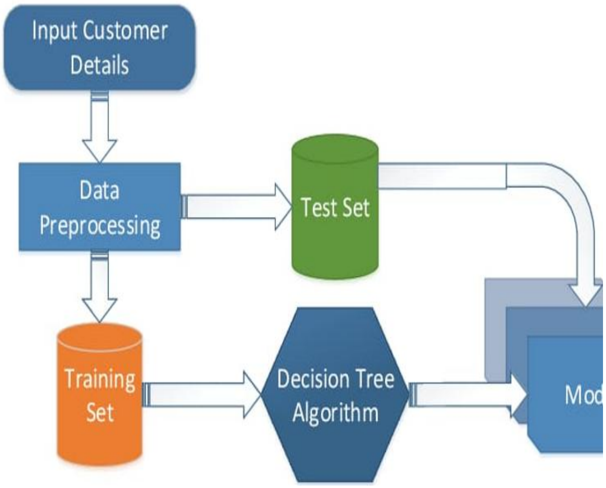
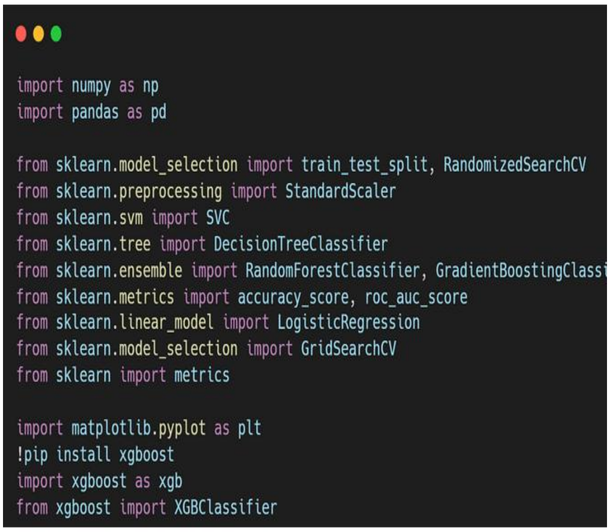


Fig.3 System architecture

## V. IMPLEMENTATION

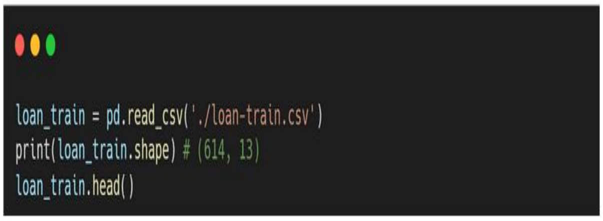
Loading Libraries for Loan Prediction using Machine Learning



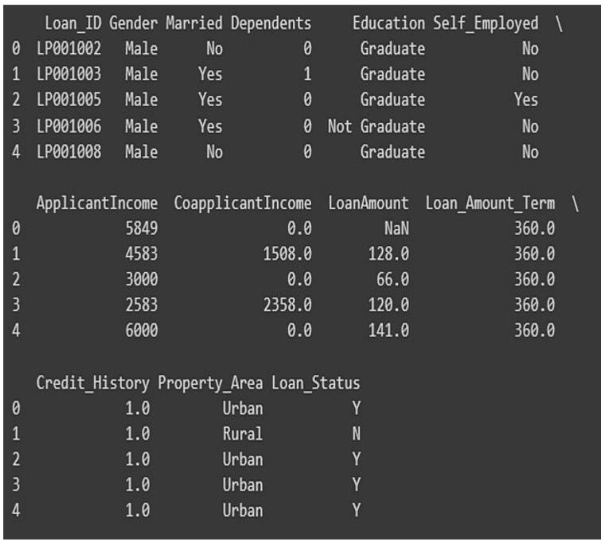
```
import numpy as np
import pandas as pd

from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassi
from sklearn.metrics import accuracy_score, roc_auc_score
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV
from sklearn import metrics

import matplotlib.pyplot as plt
!pip install xgboost
import xgboost as xgb
from xgboost import XGBClassifier
```

### Loading the Loan Prediction Dataset for the Machine Learning

Download the CSV files from the Kaggle page. Using the read_csv function from Pandas, we load the training dataset given in the competition since it is the only file with the target variable mapped.



```
loan_train = pd.read_csv('./loan-train.csv')
print(loan_train.shape) # (614, 13)
loan_train.head()
```



```
    Loan_ID Gender Married Dependents     Education Self_Employed  \
0  LP001002   Male      No          0      Graduate            No
1  LP001003   Male     Yes          1      Graduate            No
2  LP001005   Male     Yes          0      Graduate           Yes
3  LP001006   Male     Yes          0  Not Graduate            No
4  LP001008   Male      No          0      Graduate            No

   ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  \
0             5849                0.0         NaN             360.0
1             4583             1508.0       128.0             360.0
2             3000                0.0        66.0             360.0
3             2583             2358.0       120.0             360.0
4             6000                0.0       141.0             360.0

   Credit_History Property_Area Loan_Status
0             1.0         Urban           Y
1             1.0         Rural           N
2             1.0         Urban           Y
3             1.0         Urban           Y
4             1.0         Urban           Y
```

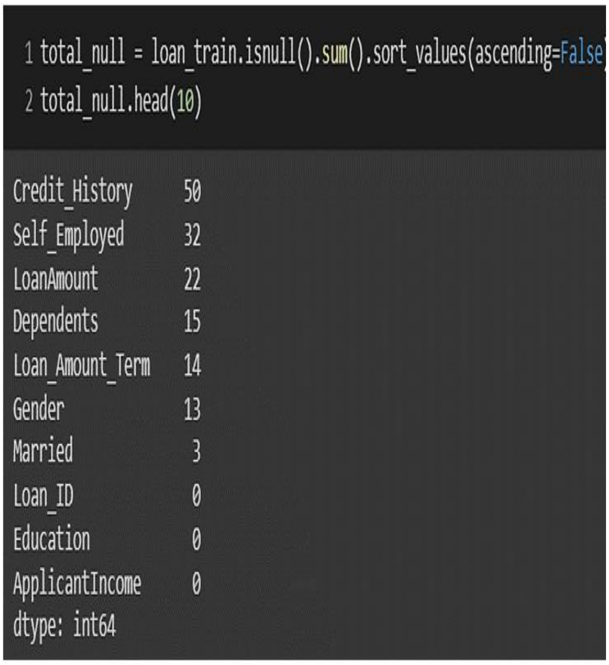
As we can see above, most variables are categorical, with most having binary categories. First, we need to encode these one-hot and then normalize the numerical variables. Finally, the Loan_ID column is not very useful since it has 614 unique values. We will remove that column from our final set.

### Data Pre-Processing for Loan Prediction using Machine Learning

Data pre-processing involves label encoding, handling missing values, selecting appropriate columns, normalization, and more. We will have to perform all of these steps on our dataset despite it being a relatively clean and structured one. Python Pandas is particularly useful in taking care of these pre-processing steps to prepare the training dataset. In-built functions of the Data Frame class can cater to EDA, cleaning, pre-processing, sorting, and filtering as needed

### Treating missing values

The isnull() method of the DataFrame class returns a binary value for every row of every column, indicating whether or not the cell is empty. Using sum() we can treat the binaries as 0 and 1 and get a count of NULL values for each column.



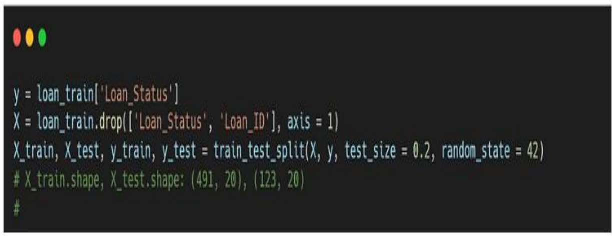
```
total_null = loan_train.isnull().sum().sort_values(ascending=False
total_null.head(10)
```

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

7 columns have a non-zero number of NULL values, with Credit_History having the most (50). Given the size of our dataset, some of these columns have many empty fields, which we cannot handle by just removing the respective rows. Doing this will significantly decrease the size of the training dataset and adversely impact the model performance. Instead, we use null value treatment methods like replacing the values with the Mean or Mode of the column values. Using mode works best in our case, as most columns are binary. Moreover, Mode will simply put the most occurring instance in place of empty fields, which, under the circumstances, would be the best guess.

### Creating Train and Test Dataset

Using the popular train_test_split function from sklearn and a split ratio of 80:20, we create the train and test data sets as follows:



```
y = loan_train['Loan_Status']
X = loan_train.drop(['Loan_Status', 'Loan_ID'], axis = 1)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
# X_train.shape, X_test.shape: (491, 20), (123, 20)
#
```

Our final train dataset has nearly 500 samples and 20 columns. We are ready to use it to train different machine learning models.

Training Machine Learning (ML) Models for Loan Prediction

We follow a fixed pipeline in trying out different models. All the models are implemented in sklearn except XGBoost. We initialize the model without any parameters and pass it through the

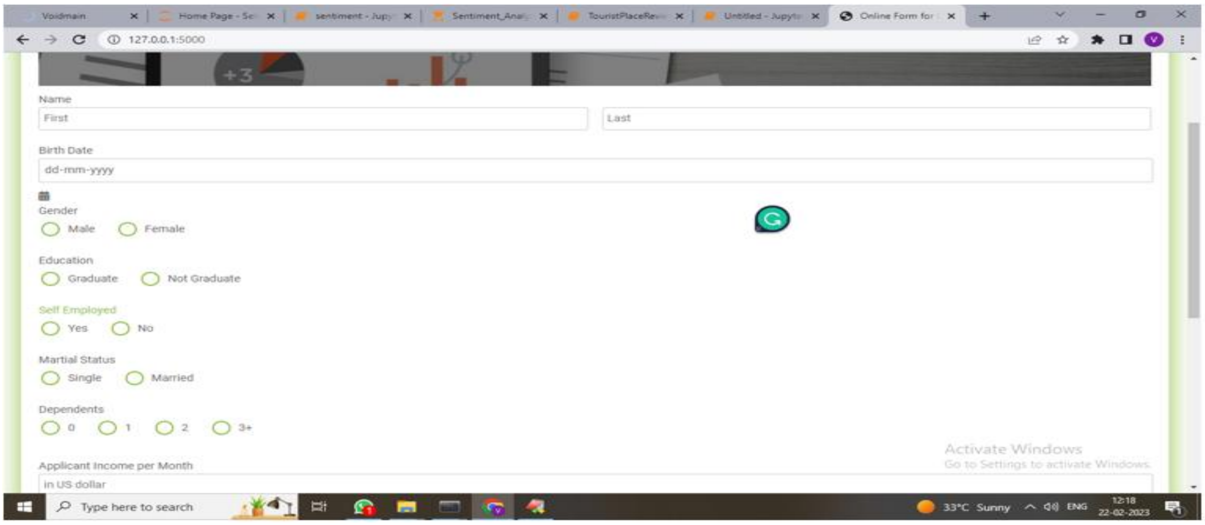
ISSN: 2366-1313

randomized cross-validation search module Randomized SearchCV with a dictionary of relevant hyperparameters. We define these hyperparameters in advance for each classification algorithm based on their availability in their sklearn implementation.

The cross-validation process will train and check the training accuracy for different permutations of these hyperparameter values and return the best-performing model. We run cross-validation for 100 iterations with a fold size of 4 samples. We will also get to see what the best choice of parameters is for every model on our training dataset.

Let's get started.

## VI. RESULTS



Name

First

Last

Birth Date

dd-mm-yyyy

Gender

○ Male ○ Female

Education

○ Graduate ○ Not Graduate

Self Employed

○ Yes ○ No

Martial Status

○ Single ○ Married

Dependents

○ 0 ○ 1 ○ 2 ○ 3+

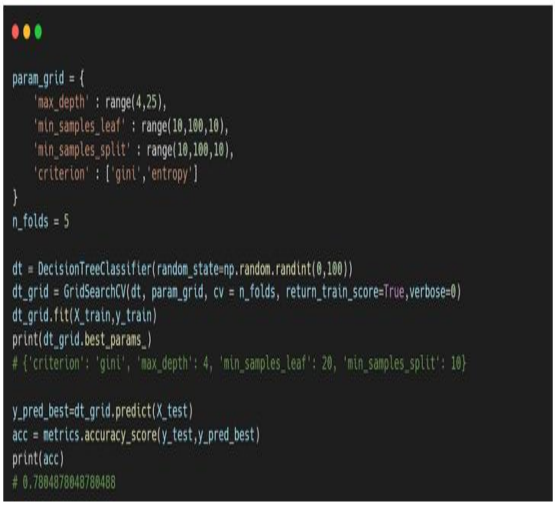
Applicant Income per Month

in US dollar

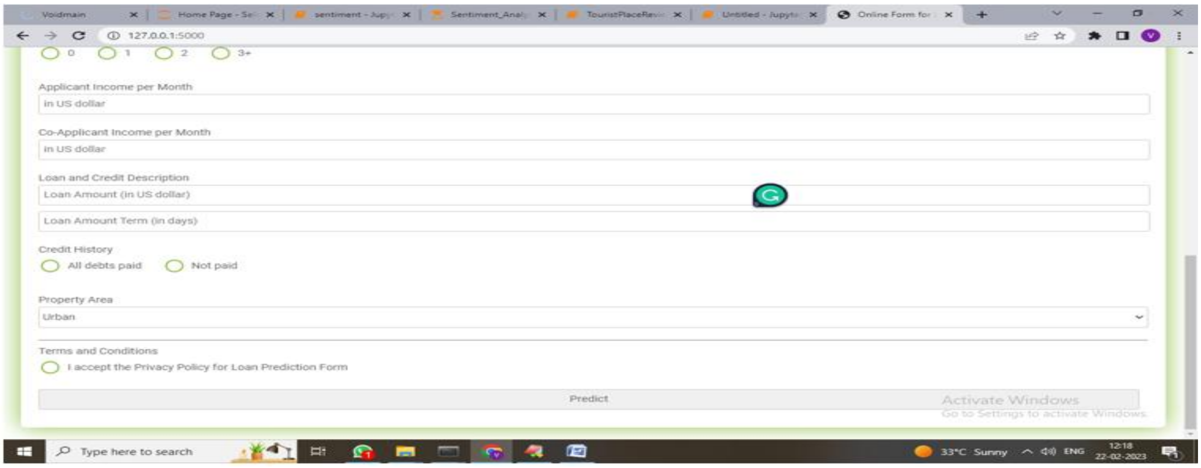
Fig.4 Give the inputs to predict loan approval

### Decision Tree

Next, we can try a single decision tree with the max depth ranging from 4 to 25 and minimum samples for leaf and split between 10 and 100. 4 is the best max depth, while the ideal criterion is the default 'Gini' index.



```
param_grid = {
    'max_depth' : range(4,25),
    'min_samples_leaf' : range(10,100,10),
    'min_samples_split' : range(10,100,10),
    'criterion' : ['gini','entropy']
}
n_folds = 5

dt = DecisionTreeClassifier(random_state=np.random.randint(0,100))
dt_grid = GridSearchCV(dt, param_grid, cv = n_folds, return_train_score=True,verbose=0)
dt_grid.fit(X_train,y_train)
print(dt_grid.best_params_)
# {'criterion': 'gini', 'max_depth': 4, 'min_samples_leaf': 20, 'min_samples_split': 10}

y_pred_best=dt_grid.predict(X_test)
acc = metrics.accuracy_score(y_test,y_pred_best)
print(acc)
# 0.7804878048780488
```



○ 0 ○ 1 ○ 2 ○ 3+

Applicant Income per Month

in US dollar

Co-Applicant Income per Month

in US dollar

Loan and Credit Description

Loan Amount (in US dollar)

Loan Amount Term (in days)

Credit History

○ All debts paid ○ Not paid

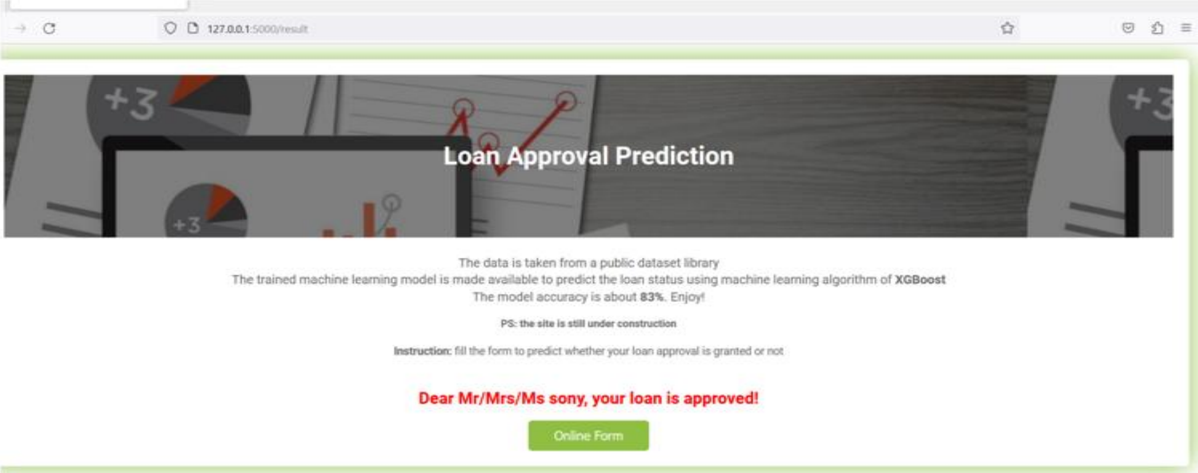
Property Area

Urban

Terms and Conditions

○ I accept the Privacy Policy for Loan Prediction Form

Fig.5 Giving the inputs to predict loan approval or reject



The data is taken from a public dataset library

The trained machine learning model is made available to predict the loan status using machine learning algorithm of **XGBoost**

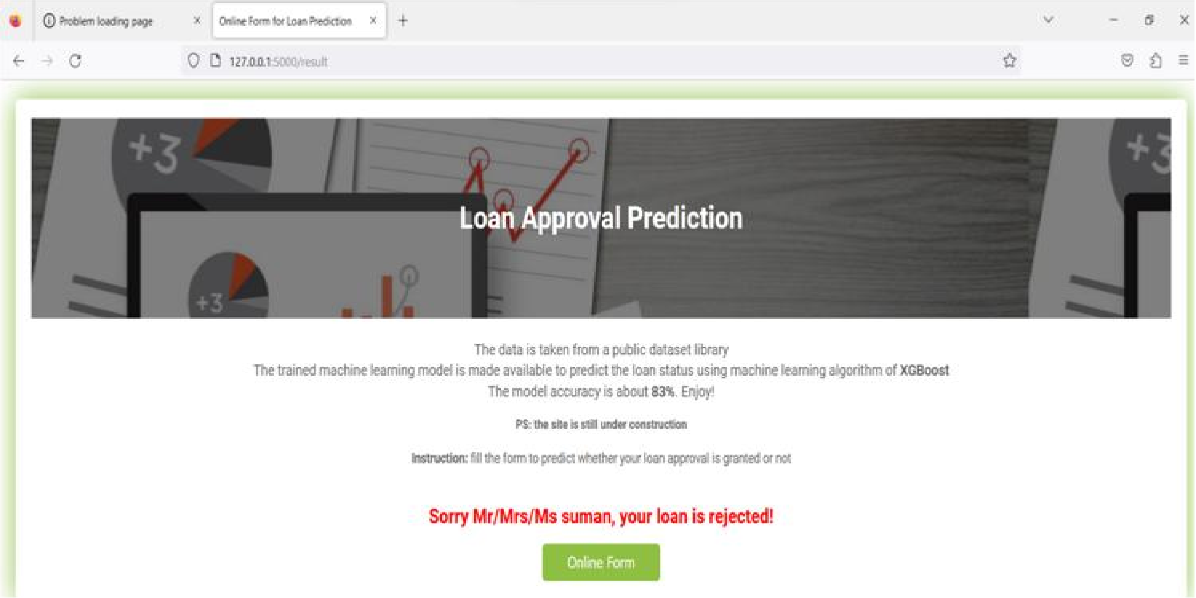
The model accuracy is about **83%**. Enjoy!

**PS: the site is still under construction**

**Instruction:** fill the form to predict whether your loan approval is granted or not

**Dear Mr/Mrs/Ms sony, your loan is approved!**

Fig.6 Loan approved



The data is taken from a public dataset library

The trained machine learning model is made available to predict the loan status using machine learning algorithm of **XGBoost**

The model accuracy is about **83%**. Enjoy!

**PS: the site is still under construction**

**Instruction:** fill the form to predict whether your loan approval is granted or not

**Sorry Mr/Mrs/Ms suman, your loan is rejected!**

Fig.7 Loan is rejected

## VII. CONCLUSION

After this work, we are able to conclude that Decision tree version is extraordinary efficient and gives a higher end result. We have developed a model which can easily predict that the person will repay its loan or not. we can see our model has reduced the efforts of bankers. Machine learning has helped a lot in developing this model which gives precise results.

## REFERENCES

[1] Bing Liu, “Sentiment Analysis: Mining Opinions, Sentiments, and Emotions,” Cambridge University Press, ISBN:978-1-107-01789-4.

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

[3] X.Frencis Jensy, V.P.Sumathi, Janani Shiva Shri, “An exploratory Data Analysis for Loan Prediction based on nature of clients”, International Journal of Recent Technology and Engineering (IJRTE),Volume-7 Issue-4S, November 2018.

[4] 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.

[5] Rattle data mining tool, http://rattle.togaware.com/rattle - download.html.

[6] Aafer Y., Du W., Yin H., Droid APIMiner: Mining API-Level Features for Robust Malware Detection in Android, Security and privacy in Communication Networks, Springer, pp 86-103, 2013.

[7] J. R. Quinlan., Induction of Decision Tree, Machine Learning, Vol. 1, No. 1. pp. 81-106

[8] W. Kluwer Bizfillings, "What Banks Look for when Reviwing a Loan Application", 2019, [online] Available: Bizfillings.com.

[9] L. Al-Blooshi and H. Nobanee, "Applications of Artificial Intelligence in Financial Management Decisions: A Mini-Review", SSRN Electronic Journal, 2020.