A MACHINE LEARNING APPLICATION FOR PREDICTION LOAN DEFAULT BASED ON CONSUMER BEHAVIOR

MINI PROJECT REPORT

Submitted by

VISHVA A 210701314
VINOTH RAJ G 210701509

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY:: CHENNAI 600 025

APRIL 2024

RAJALAKSHMI ENGINEERING COLLEGE,

CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled "**A machine learning application for predicting loan default based on consumer behavior**" is the bonafide work of **"Vishva A (210701314), Vinoth Raj G (210701509)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Karthick V

Associate Professor,
Department of Computer Science and Engineering,

Rajalakshmi Engineering College,
Chennai – 602015

Submitted to Mini Project Viva-Voce Examination held on ____________________

Internal Examiner **External Examiner**

ABSTRACT

The cost of assets is increasing day by day and the capital required to purchase an entire asset is very high. So purchasing it out of your savings is not possible. The easiest way to get the required funds is to apply for a loan. But taking a loan is a very time consuming process. The application has to go through a lot of stages and it's still not necessary that it will be approved. To decrease the approval time and to decrease the risk associated with the loan many loan prediction models were introduced. The aim of this project was to compare the various Loan Prediction Models and show which is the best one with the least amount of error and could be used by banks in the real world to predict if the loan should be approved or not taking the risk factor in mind. After comparing and analyzing the models, it was found that the prediction model based on Random Forest proved to be the most accurate and fitting of them all. This can be useful in reducing the time and manpower required to approve loans and filter out the perfect candidates for providing loans. Furthermore, feature importance analysis shows that important variables in determining the results of loan approvals include applicant income, credit history, loan amount, and loan term.By correctly identifying high-risk applicants, this loan prediction system can minimize the danger of default and ultimately increase the efficiency and profitability of lending institutions. It also has the potential to drastically reduce the human workload of loan officers. In order to improve predicted performance even further, future work will involve expanding the datasets used for model refinement, adding new features, and investigating cutting-edge methods like deep learning.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S.MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Karthiki V** Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

Vishva A - 210701314
Vinoth Raj G - 210701509

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SVM	Support Vector Machines
KNN	K Nearest Neighbours
SVC	Support Vector Classifier

CHAPTER 1
INTRODUCTION

1.1 GENERAL

The loan approval prediction project involves developing a machine learning model to predict whether a loan application will be approved based on applicant data. The process includes collecting relevant financial and personal data, preprocessing it to handle missing values and encode categorical variables, and performing exploratory data analysis to uncover patterns.

1.2 OBJECTIVE

The objective of the this project is to develop a machine learning model that accurately predicts the approval status of loan applications based on applicants' financial and personal information. By leveraging historical data, the model aims to identify patterns and key factors that influence loan approval decisions, thereby enabling financial institutions to streamline their loan processing, reduce the risk of defaults, and improve decision-making efficiency. This predictive model will serve as a tool to assist loan officers in making more informed and consistent lending decisions, ultimately enhancing customer satisfaction and operational effectiveness.

1.3 EXISTING SYSTEM

Regarding loan default prediction, current systems assess borrowers' creditworthiness mostly by conventional statistical techniques and credit scoring models. Banks, Housing Finance Companies and some NBFC deal in various types of loans like housing loan, personal loan, business loan etc in all over the part of countries. These companies have existence in Rural, Semi Urban and Urban areas. After applying for a loan by a customer , these companies validate the eligibility of customers to get the loan or not. This form consists of details like Sex, Marital Status, Qualification, Details of Dependents, Annual Income, Amount of Loan, Credit History of Applicant and others. These scores are determined by looking at things like length of credit history, new credit, quantities outstanding, payment history, and credit kinds that are employed. Credit ratings, while useful in certain situations, generally give a static picture of a borrower's creditworthiness and do not account for behavioral shifts that occur in real time. Heuristic principles and financial measures, such as loan-to-value and debt-to-income ratios, are also used in traditional loan evaluation.

1.4 PROPOSED SYSTEM

The paper will be comparing different prediction models and deduce their limitations as well as advantages. Since all the research papers used different sets of data to infer the accuracy and for cross validation of data, the authors have used the same data for all the models which will give a clearer view on their performance and lead to a better comparison of the same. On the basis of the results, a modified prediction model will be created to ensure maximum accuracy and performance.

CHAPTER 2

LITERATURE SURVEY

The author, Vaidya, Ashlesha [1] uses logistic regression as a machine learning tool in paper and shows how predictive approaches can be used in real world loan approval problems. His paper uses a statistical model (Logistic Regression) to predict whether the loan should be approved or not for a set of records of an applicant. Logistic regression can even work with power terms and nonlinear effects. Some limitations of this model are that it requires independent variables for estimation and a large sample is required for parameter estimation.

A work by Amin, Rafik Khairul and Yuliant Sibaroni [2] was referenced which used a Decision tree algorithm called C4.5 to implement a predictive model. This algorithm creates a decision tree that generally gives a high accuracy in decision making problems. Dataset of 1000 cases is used in which 70% is approved and the rest is rejected. This paper shows C4.5 algorithm performance in recognizing the eligibility of the applicant to repay his/her loan. From the conducted tests, it is found that the highest precision value is 78.08% which was found using a data partition of 90:10. The greatest recall value is 96.4% and was reached with a data partition of 80:20. Partition of 80:20 is considered to be best since it has a high recall and the highest accuracy.

The research and work done by Arora, Nisha and Pankaj Deep Kaur [3] aimed at forecasting whether an applicant can be a loan defaulter or not. It uses Bolasso to select most relevant attributes based on their robustness and then applied to classification algorithms like Random Forest, SVM, Naive Bayes and KNearest Neighbours (KNN) to test how accurately they can predict the results. It is concluded that the Bolasso enabled Random Forest algorithm (BS-RF) provides the best results in credit risk evaluation and gives better accuracy by using optimised feature selection methods.

In paper authored by Yang, Baoan, et al. [4], the use of artificial neural networks in an early warning system for predicting loan risk is discussed wherein it covers the early warning signals for deteriorating financial situations. The ability of an applicant to repay the loan is determined to be the most relevant aspect in the financial analysis. The early warning system in this paper uses an artificial neural network that is utilizing the traditional early warning concepts. This system based on ANN proves to be a very effective decision tool and early warning system for banks and other commercial lending organizations.

The scope of using Genetic Algorithms in building prediction models was also discussed in the paper by Metawa Noura, M. Kabir Hassan and Mohamed Elhoseny [5]. This paper discusses a prediction model made using Genetic Algorithm which can facilitate banks in making lending decisions in case of decrease in lending supply. The main focus of the GA model is twofold: maximizing profit and minimizing errors in loan approval in case of dynamic lending decisions. Several factors like type of loan, rating of creditor and expected loan loss are integrated to GA chromosomes and then validation is done. The result shows that GAMCC increases the profits of the bank by 3.9% to 8.1%. Yet another approach was used by Hassan, Amira Kamil Ibrahim and Ajith Abraham[6] wherein they used a German dataset and built a prediction model working basically on backpropagation and implemented with three different back propagation algorithms. They also used two different methods for two filtering functions for the attributes which resulted in DS2 giving highest accuracy using PLsFi filtering function.

CHAPTER 3
SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

Table 3.1.1 Hardware Specifications

PROCESSOR	Intel Core i5
RAM	4GB or above (DDR4 RAM)
GPU	Intel Integrated Graphics
HARD DISK	6GB
PROCESSOR FREQUENCY	1.5 GHz or above

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be preinstalled and the languages needed to develop the project has been listed out below.

Table 3.1.2 Software Specifications

BACK END	Python
SOFTWARES USED	Visual Studio, Jupyter Notebook

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

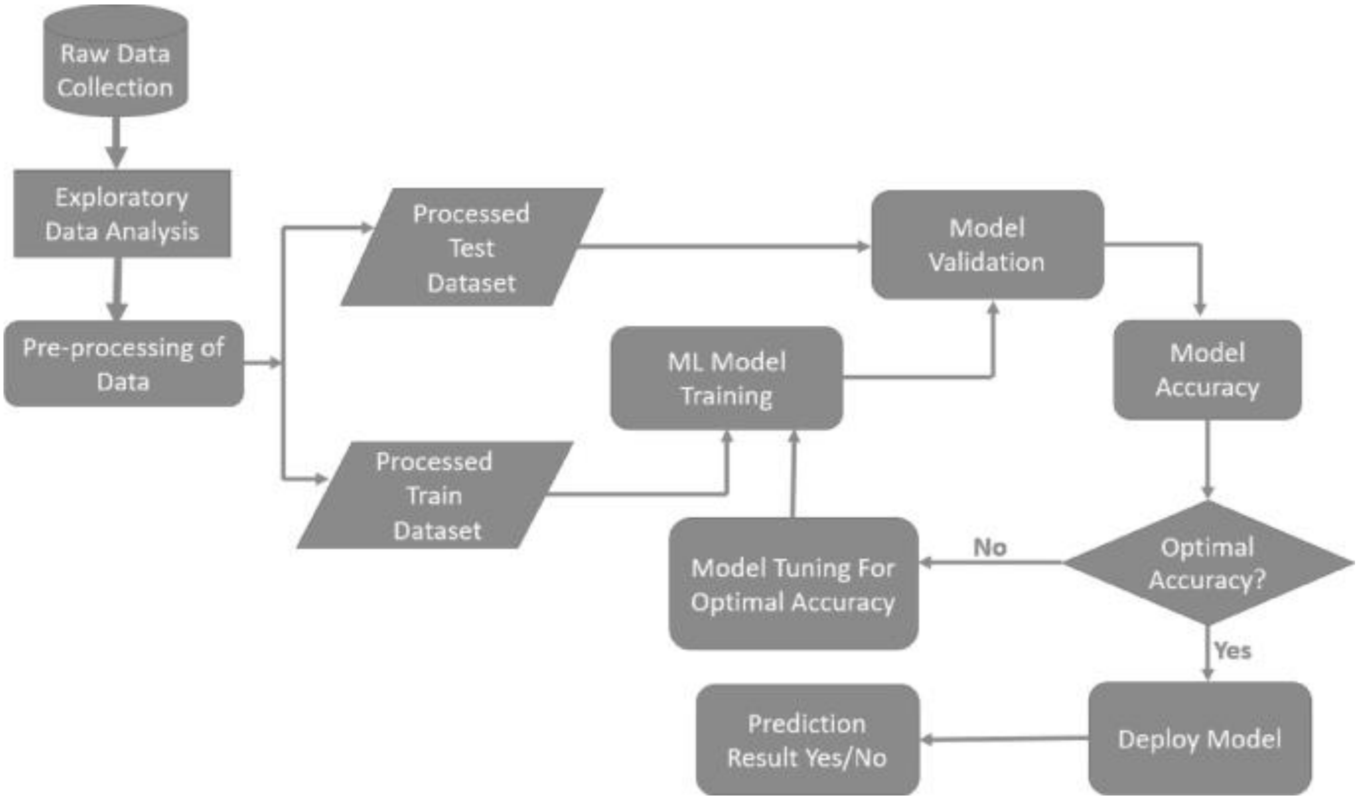


Fig 3.2.1 Architecture Diagram

CHAPTER 4
PROJECT DESCRIPTION

4.1 MODULE DESCRIPTION

4.1.1 DATA COLLECTION :

The first phase of the project involves data from trusted sources such as kaggle.The data set collected should have desired data columns and be able provide better results and the size should be sufficient enough.

4.1.2 DATA PREPROCESSING :

The Data collected won't be in a state that can be used for training purposes hence, the data should undergo the step of preprocessing in which common problems are eradicated such as missing values , improper spelling in data or incorrectness in data etc. Various python libraries specialized for data analysis can be utilized for this purpose such as Numpy, Pandas.This step is crucial for the project as these may cause inefficiency if they are fed directly to the model.

4.1.3 EDA :

EDA stands for Exploratory Data Analysis in which the entire acquired data is analyzed for its relation within the data. Any outliers or deviation of data can be inferred at this point and also this helps to gain the significance of each data column. The common libraries utilized for this step include Matplotlib and Seaborn. Both of these are visualization tools commonly used in the project. Through EDA, we concluded that several attributes of users such as phone number, user id etc. are redundant and thus they are dropped. Heatmaps are extensively used to know the correlation between various attributes.

4.1.4 MODEL TRAINING :

The vectorized text data is used to train a convolutional neural network model . During training, the model adjusts its internal parameters iteratively to minimize a defined loss function. Dropout layers are included to prevent overfitting, ensuring the model generalizes well to unseen data. The model is trained using a portion of the data, while performance is monitored using a separate validation set.

4.1.5 MODEL EVALUATION :

Once training is complete, the model's performance is evaluated using a separate test dataset. Performance metrics such as accuracy, precision, and recall are calculated to assess the model's effectiveness in classifying legal descriptions.

CHAPTER 5

IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION

5.1.1 Importing libraries and dataset

Firstly we have to import libraries :

- Pandas – Python library used to load the Data Frame
- Matplotlib – Python library visualize the data features i.e. barplot
- Seaborn – Python library to see the correlation between features using heatmap

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

data = pd.read_csv("LoanApprovalPrediction.csv")
```

```
C:\Users\admin\AppData\Local\Temp\ipykernel_12880\3117320796.py:1: DeprecationWarning:
Pyarrow will become a required dependency of pandas in the next major release of pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and better interoperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54466

  import pandas as pd
```

After importing our dataset, let's view it by using a simple method,

```
data.head(5)
```

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

Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
360.0	1.0	Urban	Y
360.0	1.0	Rural	N
360.0	1.0	Urban	Y
360.0	1.0	Urban	Y
360.0	1.0	Urban	Y

5.1.2 Data Preprocessing and Visualization

In this step, we get the number of columns of object data type.

```
obj = (data.dtypes == 'object')
print("Categorical variables:",len(list(obj[obj].index)))
```

```
Categorical variables: 7
```

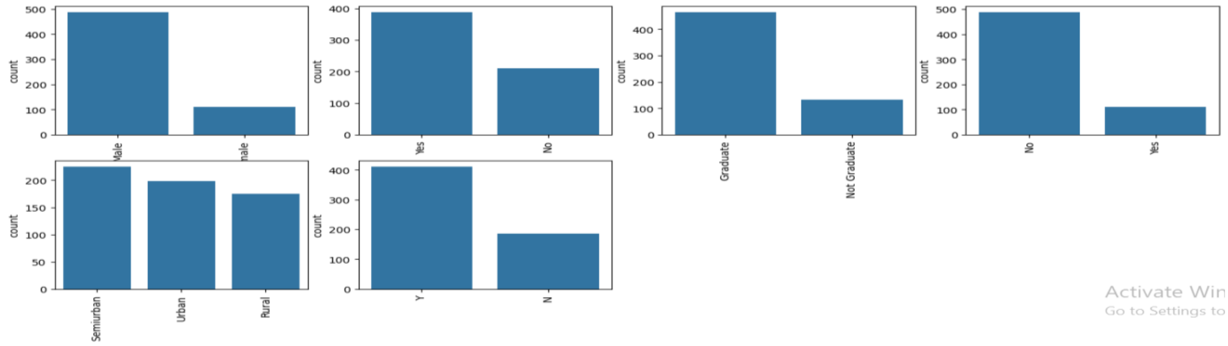
Then, as Loan_ID is completely unique and it is not correlated with any of the other columns, we drop it using the drop() function.

```
# Dropping Loan_ID column
data.drop(['Loan_ID'],axis=1,inplace=True)
```

5.1.3 Visualizing all the unique values in columns using barplot will simply show which value is dominating as per our dataset.

```
obj = (data.dtypes == 'object')
object_cols = list(obj[obj].index)
plt.figure(figsize=(18,36))
index = 1

for col in object_cols:
    y = data[col].value_counts()
    plt.subplot(11,4,index)
    plt.xticks(rotation=90)
    sns.barplot(x=list(y.index), y=y)
    index +=1
```



As we see, As all the categorical values are binary so we can use Label Encoder for all such columns and the values will change into **int** datatype.

```
# Import label encoder
from sklearn import preprocessing

# label_encoder object knows how
# to understand word labels.
label_encoder = preprocessing.LabelEncoder()
obj = (data.dtypes == 'object')
for col in list(obj[obj].index):
    data[col] = label_encoder.fit_transform(data[col])
```

Again we check for the object datatype columns finding out if there is still any left.

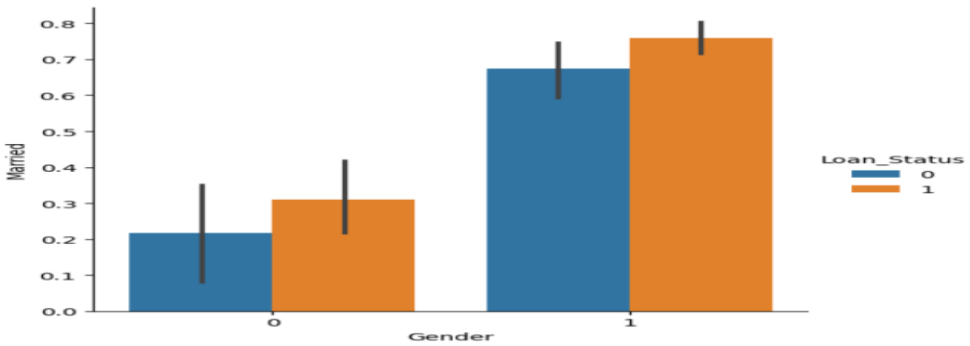
```
# To find the number of columns with
# datatype==object
obj = (data.dtypes == 'object')
print("Categorical variables:",len(list(obj[obj].index)))
```

```
Categorical variables: 0
```

Using Catplot, we visualize the plot for the Gender, and Marital Status of the applicant.

```
sns.catplot(x="Gender", y="Married",
            hue="Loan_Status",
            kind="bar",
            data=data)
```

```
<seaborn.axisgrid.FacetGrid at 0x28b6997a0c0>
```



5.2 OUTPUT SCREENSHOTS

As this is a classification problem, we will be using the models like

- KNeighborsClassifiers
- RandomForestClassifiers
- Support Vector Classifiers (SVC)
- Logistics Regression

We will use the accuracy score function from scikit-learn library to predict the accuracy We will use the accuracy score function from scikit-learn library to predict the accuracy .

Table 5.2.1 Accuracy Scores

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression

from sklearn import metrics

knn = KNeighborsClassifier(n_neighbors=3)
rfc = RandomForestClassifier(n_estimators = 7,
                             criterion = 'entropy',
                             random_state =7)
svc = SVC()
lc = LogisticRegression()

# making predictions on the training set
for clf in (rfc, knn, svc,lc):
    clf.fit(X_train, Y_train)
    Y_pred = clf.predict(X_train)
    print("Accuracy score of ",
        clf.__class__.__name__,
        "=",100*metrics.accuracy_score(Y_train,
                                       Y_pred))
```

```
Accuracy score of  RandomForestClassifier = 98.04469273743017
Accuracy score of  KNeighborsClassifier = 78.49162011173185
Accuracy score of  SVC = 68.71508379888269
Accuracy score of  LogisticRegression = 79.88826815642457
```

Prediction of test set

```
# making predictions on the testing set
for clf in (rfc, knn, svc,lc):
    clf.fit(X_train, Y_train)
    Y_pred = clf.predict(X_test)
    print("Accuracy score of ",
        clf.__class__.__name__,"=",
        100*metrics.accuracy_score(Y_test,
                                   Y_pred))
```

```
Accuracy score of  RandomForestClassifier = 82.5
Accuracy score of  KNeighborsClassifier = 63.74999999999999
Accuracy score of  SVC = 69.16666666666667
Accuracy score of  LogisticRegression = 80.0
```

```
plt.figure(figsize=(12,6))

sns.heatmap(data.corr(),cmap='BrBG',fmt='.2f',
            linewidths=2,annot=True)
```

<Axes: >

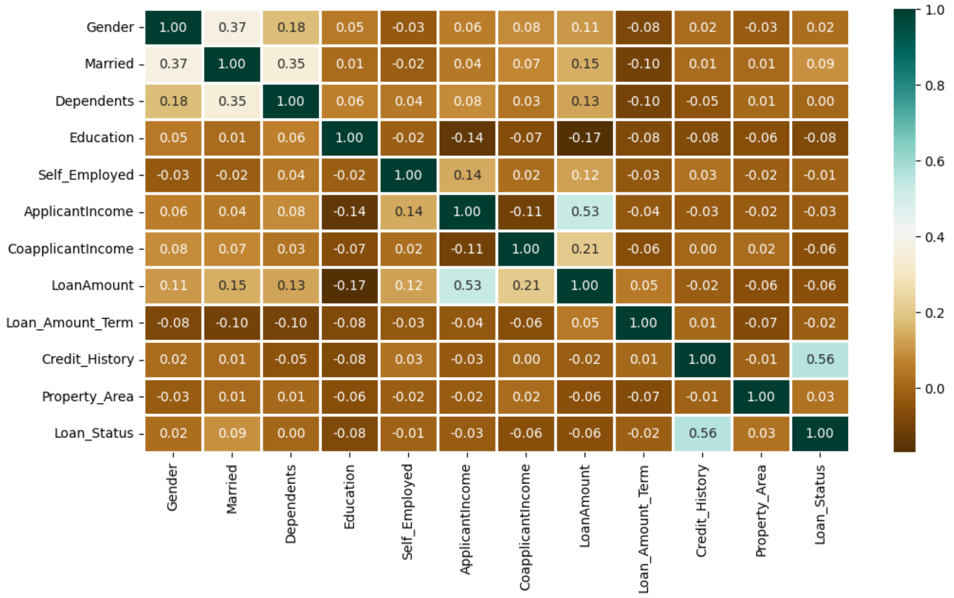


Fig 5.2.1 Confusion Matrix

The above heatmap is showing the correlation between Loan Amount and ApplicantIncome. It also shows that Credit_History has a high impact on Loan_Status.

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENTS

6.1 CONCLUSION

The implementation of the system to analyze the allocation of loan to the individual based on their details of the past and current details with the help of machine learning algorithm provides insights to the banking institutes to set the basic criteria for allocation of loan to them. The accuracy of the algorithm is useful for setting the criteria for each of the predictions for individuals. The RandomForestClassifier algorithm provides the overall highest accuracy with the dataset compared to the other algorithms implemented. Each algorithm has its own merits and demerits. The algorithm with highest accuracy is used.

6.2 FUTURE ENHANCEMENTS

The system can be improvised with the implementation of the updated algorithms with the help of real time data updation and training the model in a periodic way can improvise the accuracy. Providing a dataset which consists of present criteria for the loan prediction will improvise the algorithm's accuracy and the implementation of the system will be made efficient.

REFERENCES

[1] Vaidya and Ashlesha, Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval, 2021 13th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2021.

[2] Amin, Rafik Khairul and Yuliant Sibaroni, Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasargad of Yogyakarta Special Region), 2022 8rd International Conference on Information and Communication Technology (ICoICT). IEEE, 2022.

[3] Arora, Nisha and Pankaj Deep Kaur, A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment, Applied Soft Computing 86 (2020), 105936.

[4] Yang, Baoan, et al, An early warning system for loan risk assessment using artificial neural networks, Knowledge-Based Systems 14.5-6 (2024), 303-306.

[5] Metawa, Noura, M. Kabir Hassan and Mohamed Elhoseny, Genetic algorithm based model for optimizing bank lending decisions, Expert Systems with Applications 80 (2022), 75-82.

[6] Hassan, Amira Kamil Ibrahim and Ajith Abraham. "Modeling consumer loan default prediction using ensemble neural networks, 2020 International Conference On Computing, Electrical And Electronic Engineering (ICCEEE). IEEE, 2020.

[7] 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.

[8] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash,“Loan Prediction by using Machine Learning Models”, International Journal of Engineering and Techniques.Volume 5 Issue 2, Mar-Apr 2023

[9] Kumar Arun, Garg Ishan, Kaur Sanmeet, ―Loan Approval Prediction based on Machine Learning Approach‖, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 18, Issue 3, pp. 79-81, Ver. I (May-June. 2021).

[10] Aboobyda Jafar Hamid and Tarig Mohammed Ahmed, ―Developing Prediction Model of Loan Risk in Banks using Data Mining‖, Machine Learning and Applications: An International Journal (MLAIJ), Vol.3, No.1, pp. 1-9, March 2023.

[11] S. Vimala, K.C. Sharmili, ―Prediction of Loan Risk using NB and Support Vector Machine‖, International Conference on Advancements in Computing Technologies (ICACT 2022), vol. 4, no. 2, pp. 110-113, 2022.

[12] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma, Namburi Vimala Kumari, kVikash,"Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques.Volume 5 Issue 2, Mar-Apr 2023

[13] Nikhil Madane, Siddharth Nanda,"Loan Prediction using Decision tree", Journal of the Gujarat Research History, Volume 21 Issue 14s, December 2022.