LOAN ELIGIBILITY PREDICTION

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

by

L.SRIVARDHAN RAO	**(2103A51054)**
T.ABHILASH RAO	**(2103A51248)**
D.HRITHVIK REDDY	**(2103A51313)**

Under the guidance of

Mr. D. RAMESH

Assistant Professor, Department of CSE.

Department of Computer Science and Artificial Intelligence

SR UNIVERSITY

Department of Computer Science and Artificial Intelligence

CERTIFICATE

This is to certify that project entitled **"LOAN ELIGIBILITY PREDICTION**" is the bonafied work carried out by **L.SRIVARDHAN RAO, T.ABHILASH RAO, D.HRITHVIK REDDY** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2022-2023 under our guidance and Supervision.

Mr. D. RAMESH
Asst. Professor,
S R University,
Ananthasagar ,Warangal

Dr. M. Sheshikala
Assoc. Prof .& HOD (CSE)
S R University,
Ananthasagar ,Warangal

ACKNOWLEDGEMENT

We express our thanks to Course co-coordinator **Mr. D.Ramesh, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude to Head of the department CS&AI, **Dr. M.Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

ABSTRACT

In this modern world, financial institutions are playing a very crucial role. Nowadays, banks are developing their financial reserves by providing different kinds of loans to people who are in need. At the same time, there is also a massive increase in the count of individuals requesting loans. However, banks cannot provide loans for everyone as there are only limited reserves associated with each of them. So, banks must follow some stringent verification process to approve the 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 reserves of the bank and also onto the banking sector. So, banks started to 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 typical and risky process. In this project, we will develop a model to predict who is eligible for a loan in order to reduce the risk associated with the decision process and to modify the typical loan approval process into a much easier one. Moreover, we will make use of previous data of loan decisions made by the company and with the help of various data mining techniques, we will develop a loan approval decision predicting model which can draw decisions for each individual based on the information provided by them. We will use a machine-learning-based KNN, Decision-tree, Naïve Bayes algorithms to train the model. This project primary goal is to develop a loan prediction model with a better accuracy rate.

Table of Contents

1.INTRODUCTION

1.1 OVERVIEW

Distribution of the loans is the core business part of almost every banks. The main portion the bank's assets is directly came from the profit earned from the loans distributed by the banks. The prime objective in banking environmentisto invest their assets in safe hands where it is. Today many banks/financial companies approves loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning technique. The disadvantage of this model is that it emphasize different weights to each factor but in real life sometime loan can be approved on the basis of single strong factor only, which is not possible through this system.

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantagesto the bank. The Loan Prediction System can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight .A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan Prediction System allows jumping to specific application so that it can be check on priority basis. This Paper is exclusively for the managing authority of Bank/finance company, whole process of prediction is done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various department of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities.

1.2 PROBLEM STATEMENT

To design and implement the system using machine learning and data mining to predict the probability of the user to get loan or not from bank to improve the accuracy and to minimize the frauds. 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 loan by customer these companies validates the eligibility of customers to get the loan or not. This paper provides a solution to automate this process by employing machine learning algorithm. So the customer will fill an online loan application form. This form consist details like Sex, Marital Status, Qualification, Details of Dependents, Annual Income, Amount of Loan, Credit History of Applicant and others. To automate this process by using machine learning algorithm, First the algorithm will identify those segments of the customers who are eligible to get loan amounts so bank can focus on these customers.

1.3 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. The artificial neural network model for predict the credit risk of a bank. The Feed- forward back propagation neural network is used to forecast the credit default. The method in which two or more classifiers are combined together to produce a ensemble model for the better prediction. They used the bagging and boosting techniques and then used random forest technique. The process of classifiers is to improve the performance of the data and it gives better efficiency. In this work, the authors describe various ensemble techniques for binary classification and also for multi class classification. The new technique that is described by the authors for ensemble is COB which gives effective performance of classification but it also compromised with noise and outlier data of classification. Finally they concluded that the ensemble based algorithm improves the results for training data set.

Drawback of Existing System: 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.

1.4 PROPOSED SYSTEM

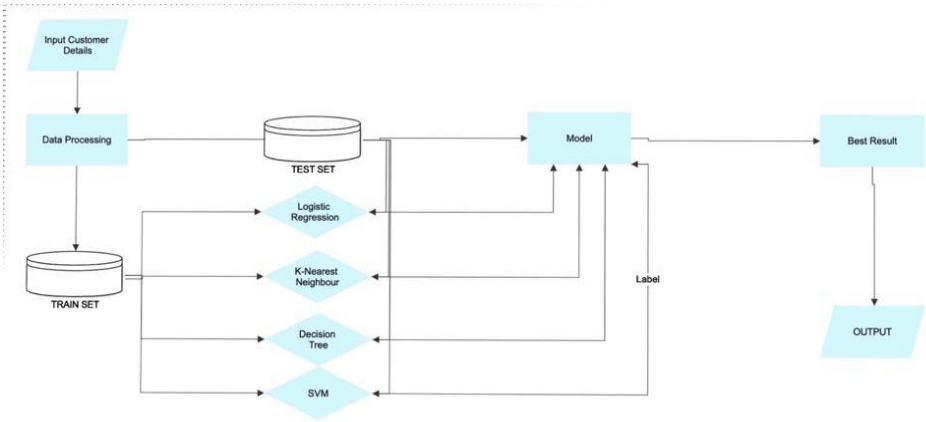
To deal with the problem, we developed automatic loan prediction using machine learning techniques. We will train the machine with previous dataset. So machine can analyse and understand the process . Then machine will check for eligible applicant and give us result.

1.5 OBJECTIVES

- Time period for loan sanctioning will be reduced.
- Whole process will be automated , so human error will be avoided .
- Eligible applicant will be sanctioned loan without any delay.
- Work burden on employees is reduced.
- If used good algorithm we can improve accuracy.
- Time saving for employees.
- No loss for money lending companies.

1.6 ARCHITECTURE

The training data set is now supplied to machine learning model, on the basis of this data set the model is trained. Every new applicant detail filled at the time of application form acts as a test data set. After the operation of testing, model predict whether the new applicant is a fit case for approval of the loan or not based upon the inference it concludes on the basis of the training data sets.



2.LITERATURE SURVEY

2.1.1 SURVEY ON PROBLEM STATEMENT

In paper [1] researchers used various machine learning algorithms to develop a machine learning model. They designed an individual model for every machine learning algorithm used in developing a protection model. They made use of different machine algorithms like International Journal of Engineering in Computer Science http://www.computersciencejournals.com/ ~ 33 ~ decision tree algorithm, Bayes algorithm, random forest algorithm. Among all the models the model built with the decision tree algorithm has scored the highest accuracy rate of 81%. It is followed by random forest algorithm with an accuracy rate of 77% and the Bayes algorithm with an accuracy rate of 69%.

In paper [2] researchers used k means clustering technique for developing a loan prediction model based on risk percentage. This model works on the principle of risk associated with each loan applicant if the model predicted that the applicant is having a low-risk percentage which means that he/she can pay back the loan, his/her loan will be approved. By using K means clustering they divided all the observation in the data into 5 clusters where each cluster is associated with some risk percentage. By using this model, they have scored an accuracy rate of 84.56%.

In paper [3] researchers used different machine learning algorithms like KNN, Decision tee, Random forest, SVM for building a model for accuracy predictor loan risk. This model deals with predicting the risk associated with each applicant which means the possible chance of loan repayment by a customer. In this paper [3] they used R language to predict the accuracy of several models. Here random forest has scored the highest possible accuracy of 82.64% in run 3 and it is followed by an accuracy rate of 81.94% using SVM in the run 3.

3.DATA PRE-PROCESSING

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessingis required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. It is the first and crucial stepwhile creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data and while doing any operation with data. It is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task. In this particular section we re-label and convert some categorical features into numericvalues. This is crucial for training machine learning models since machine learning models accepts the numeric values.

3.1 DATA DESCRIPTION

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	(Target) Loan approved (Y/N)

3.2 DATA CLEANING

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

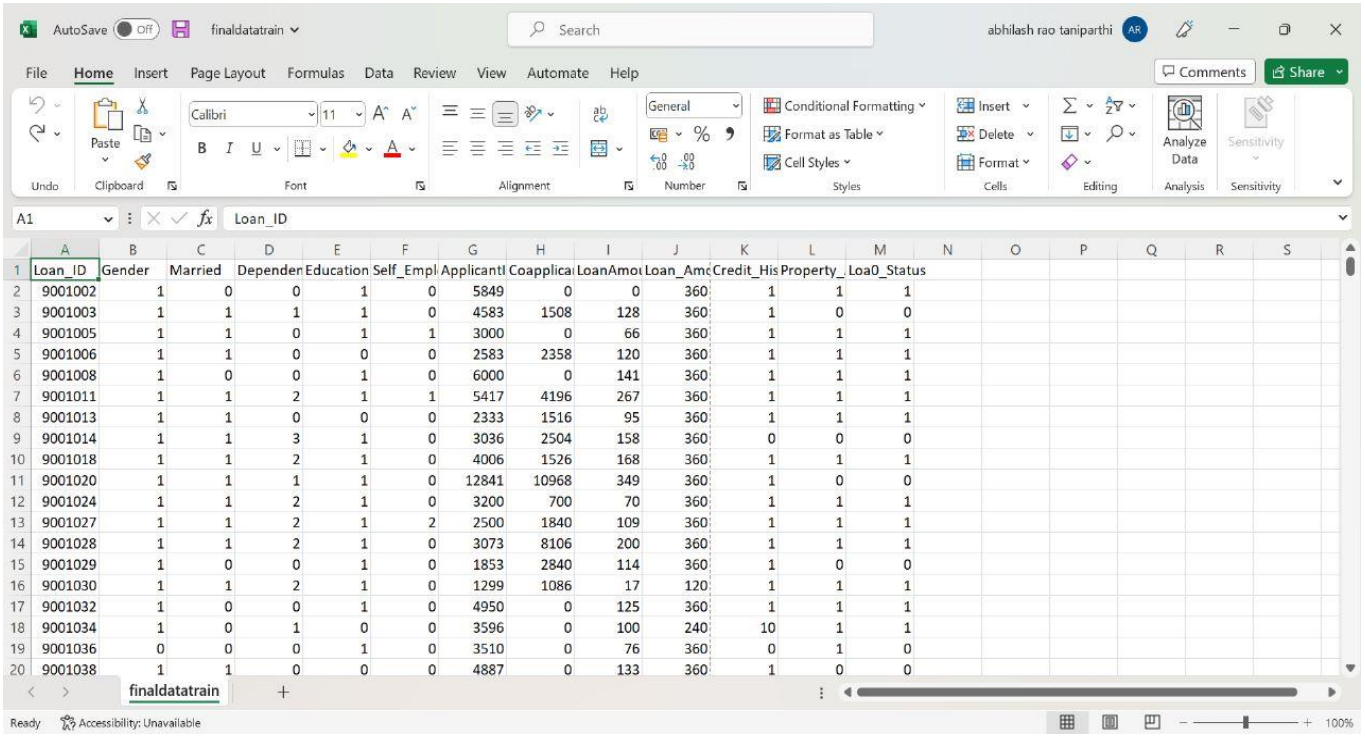
3.3 DATA AUGMENTATION

Data augmentation is a technique used to increase the diversity of a dataset by randomly creating new slightly modified versions of the instances that are already present in the dataset. It is not limited only to image data, however it is probably the easiest to understand with it. Data augmentation allows to have better data quality in the sense that the data is richer and may give better performance in a machine learning experiment. With all those methods, we can create a few more versions of each image in a dataset, with random parameters of shift and rotation. We decide to create 8 new pictures from 1 picture. This number can be changed as the data scientist wishes.

3.4 DATA VISUALISATION

```
import pandas as pd
import matplotlib.pyplot as pt
d=pd.read_csv("/content/finaldatatrain.csv")
print(d)
```

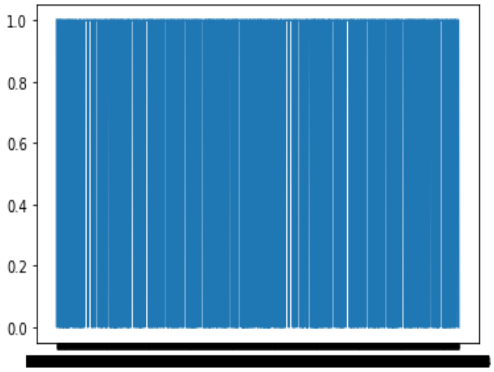
Loan_ID	Gender	Married	Dependen	Education	Self_Empl	ApplicantI	Coapplicar	LoanAmou	Loan_Amc	Credit_His	Property_A	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	1
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	0
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	1
LP001006	Male	Yes	0	Not Gradu	No	2583	2358	120	360	1	Urban	1
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	1
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	1
LP001013	Male	Yes	0	Not Gradu	No	2333	1516	95	360	1	Urban	1
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	0
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	1
LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	0
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	1
LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	1
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	1
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	0
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	1
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	1
LP001034	Male	No	1	Not Gradu	No	3596	0	100	240		Urban	1
LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	0
LP001038	Male	Yes	0	Not Gradu	No	4887	0	133	360	1	Rural	0
LP001041	Male	Yes	0	Graduate		2600	3500	115		1	Urban	1
LP001043	Male	Yes	0	Not Gradu	No	7660	0	104	360	0	Urban	0
LP001046	Male	Yes	1	Graduate	No	5955	5625	315	360	1	Urban	1
LP001047	Male	Yes	0	Not Gradu	No	2600	1911	116	360	0	Semiurban	0
LP001050		Yes	2	Not Gradu	No	3365	1917	112	360	0	Rural	0
LP001052	Male	Yes	1	Graduate		3717	2925	151	360		Semiurban	0



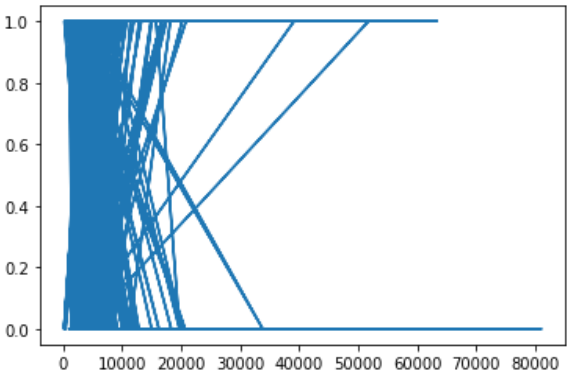
Loan_ID	Gender	Married	Dependen	Education	Self_Empl	ApplicantI	Coapplica	LoanAmou	Loan_Amc	Credit_His	Property_	Loa0_Status
9001002	1	0	0	1	0	5849	0	0	360	1	1	1
9001003	1	1	1	1	0	4583	1508	128	360	1	0	0
9001005	1	1	0	1	1	3000	0	66	360	1	1	1
9001006	1	1	0	0	0	2583	2358	120	360	1	1	1
9001008	1	0	0	1	0	6000	0	141	360	1	1	1
9001011	1	1	2	1	1	5417	4196	267	360	1	1	1
9001013	1	1	0	0	0	2333	1516	95	360	1	1	1
9001014	1	1	3	1	0	3036	2504	158	360	0	0	0
9001018	1	1	2	1	0	4006	1526	168	360	1	1	1
9001020	1	1	1	1	0	12841	10968	349	360	1	0	0
9001024	1	1	2	1	0	3200	700	70	360	1	1	1
9001027	1	1	2	1	2	2500	1840	109	360	1	1	1
9001028	1	1	2	1	0	3073	8106	200	360	1	1	1
9001029	1	0	0	1	0	1853	2840	114	360	1	0	0
9001030	1	1	2	1	0	1299	1086	17	120	1	1	1
9001032	1	0	0	1	0	4950	0	125	360	1	1	1
9001034	1	0	1	0	0	3596	0	100	240	10	1	1
9001036	0	0	0	1	0	3510	0	76	360	0	1	0
9001038	1	1	0	0	0	4887	0	133	360	1	0	0

GRAPHS PLOTTED BETWWEN FEATURE AND TARGET VARIABLES:

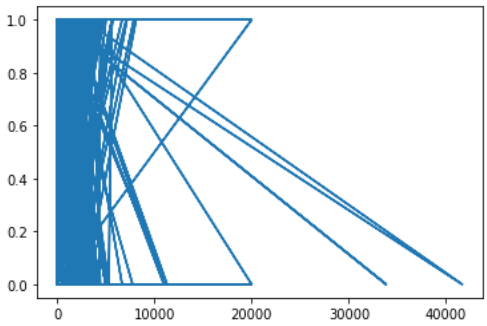
LOAN ID



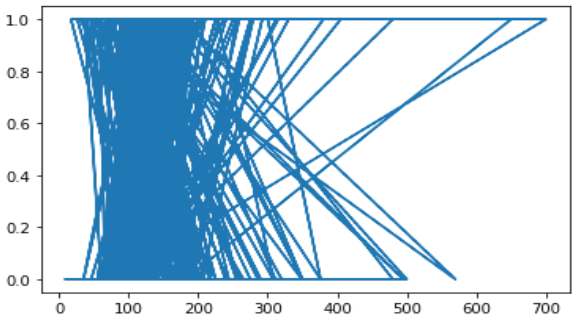
APPLICANT INCOME



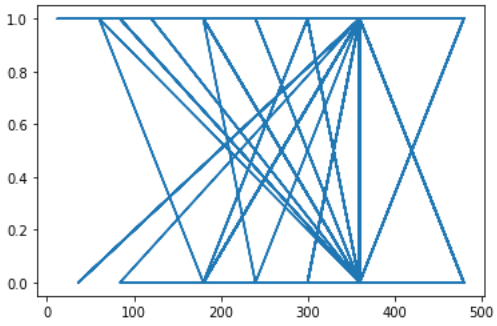
CO APPLICANT INCOME



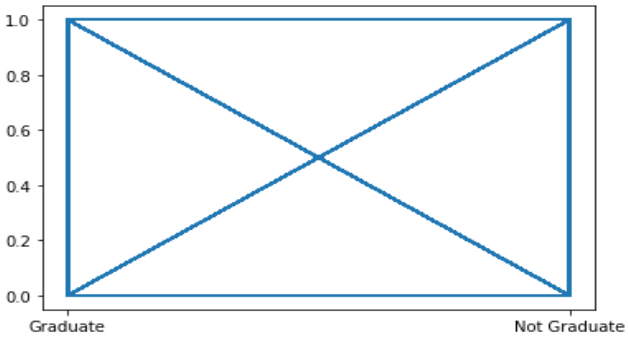
LOANAMOUNT



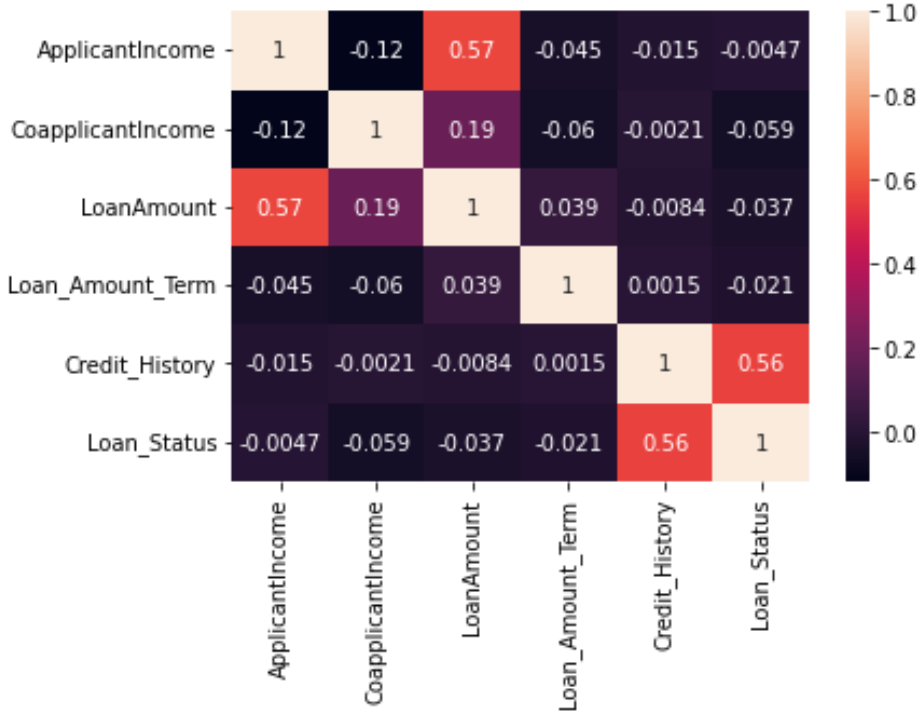
LOAN AMOUNT TEAM



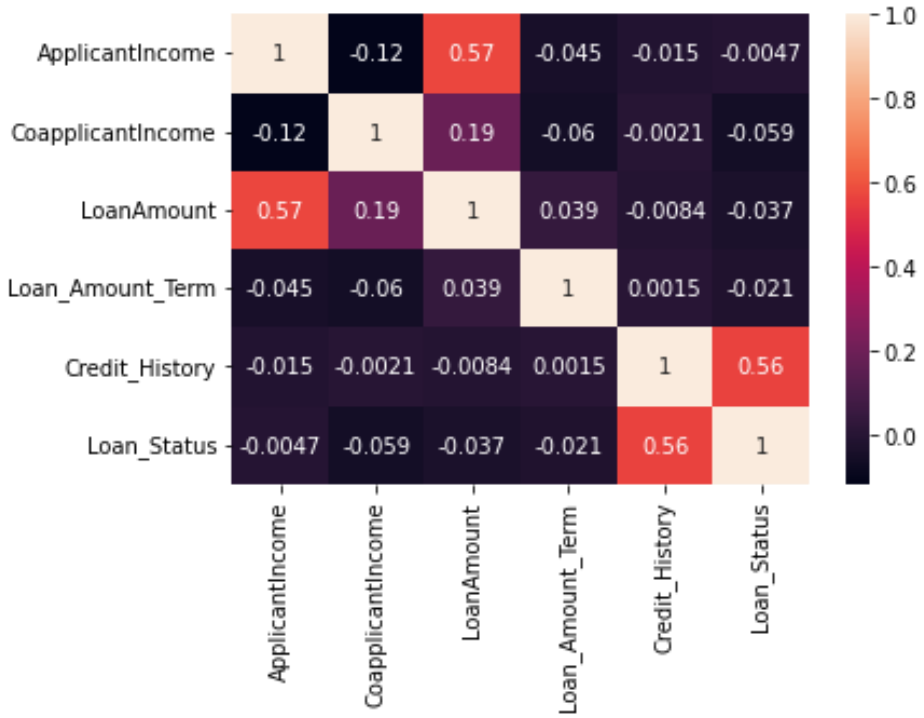
EDUCATION



Correlation Matrix:



Covariance Matrix:



4.METHODOLOGY

4.1 Logistic Regression :

Logistic regression is a statistical method used for binary classificaton problems it works by fitting logistic functions to the input variables,which transfers the input variables into a range between 0 and 1.The output of a logistic regression is binary value.

Accuracy value for logistic regression is:

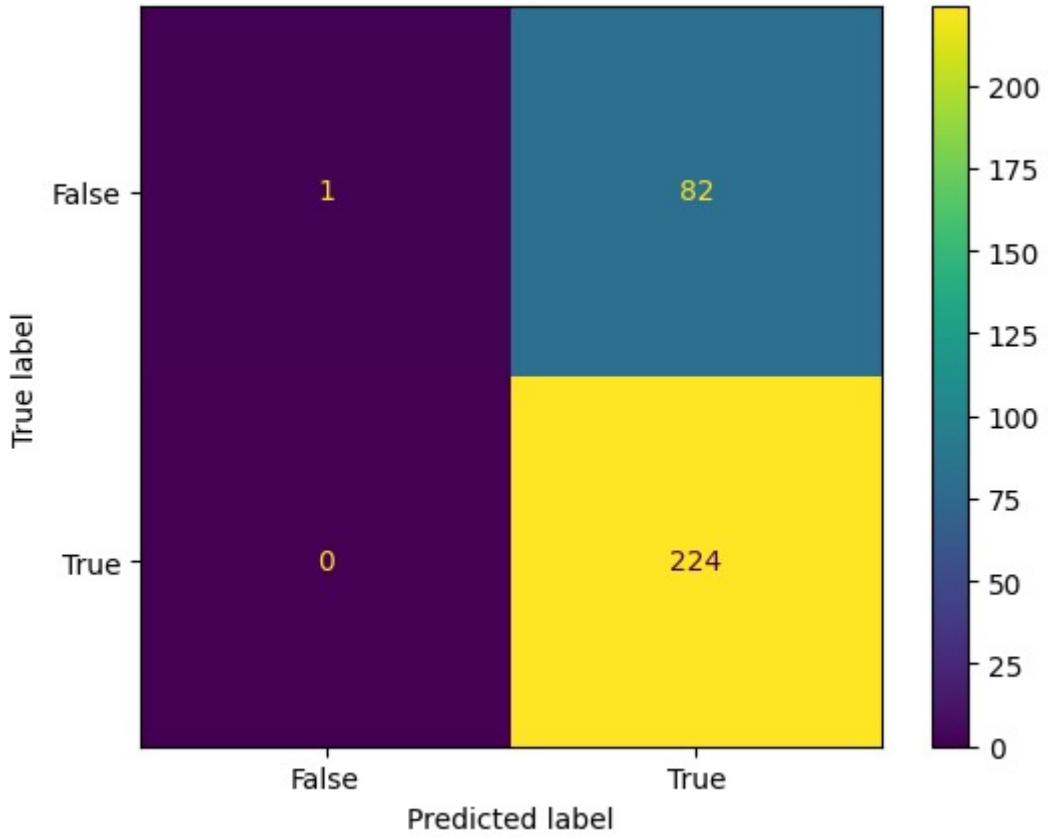
print('Logistic Regression accuracy = ',metrics.accuracy_score(lr_prediction,y_test))

print("y_predicted",lr_prediction)

print("y_test",y_test)

Logistic Regression accuracy = 0.7328990228013029

Graphical Representation of Confusion matrix :



4.2 Decision Trees :

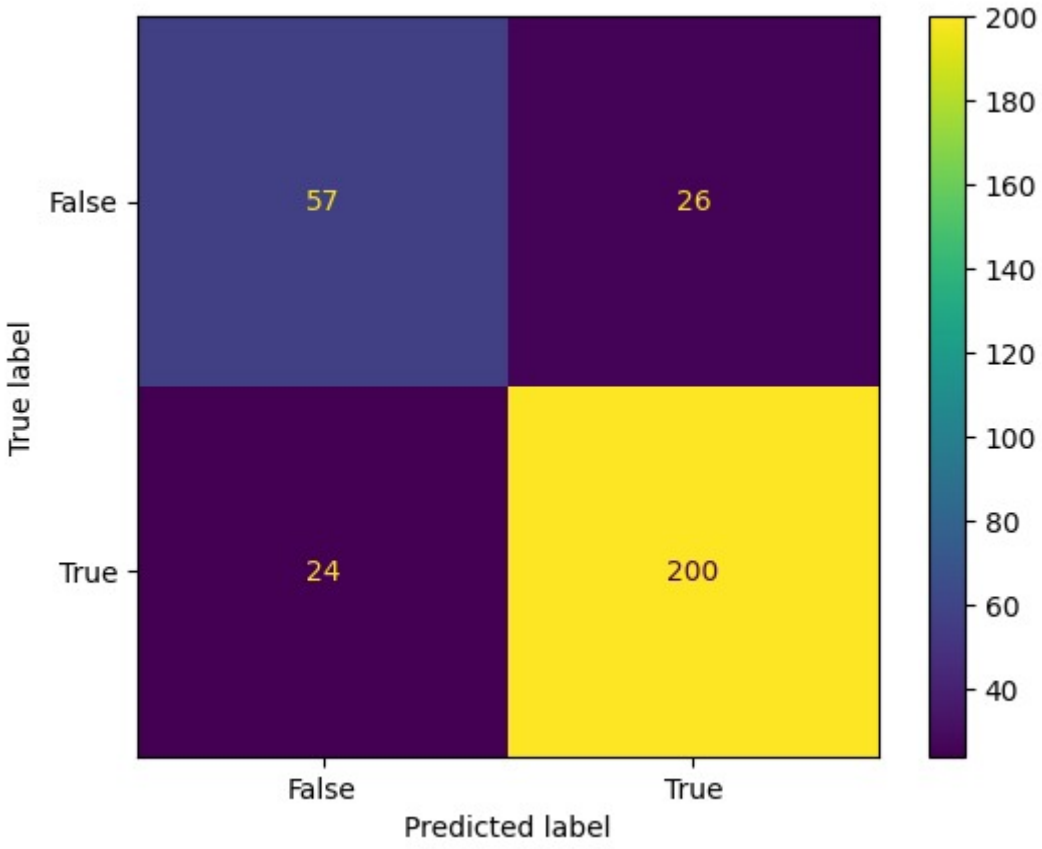
A decision tree algorithm that is used for both classification and regression problem it works by recursively partitioning the data into subsets based on the values of input features the algorithm starts by selecting the best feature to split the data into groups once the tree is bulit new instances can be classified by traversing the tree from the root node to leaf node based on the values of input.

Accuracy for decision tree:

print(accuracy_score(dp,y_test))

0.8371335504885994

Graphical Representation of Confusion matrix :



4.3 KNN :

The k nearest neighbours(K-NN) algorithm is a type of supervised machine learning it is used for both classification and regression.

The KNN algorithm works by finding the k nearest data points to a given test point uses the class label to predict the class of the test point it is used to measure similarity between data points by euclidean distance

Accuracy for KNN:

new_data = np.array([[0.0, 0.0, 0.0, 1.0, 0.0, 5000, 0.0, 50000, 240, 0.0, 2]])

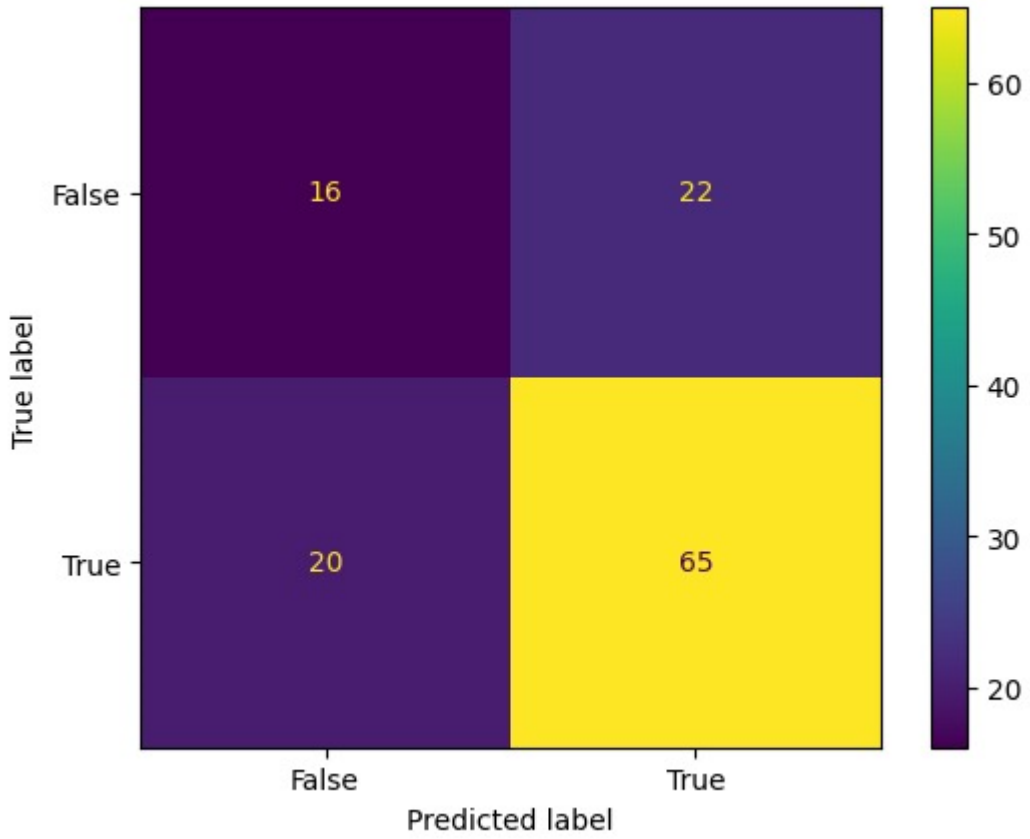
knn.predict(new_data)

Score for the training set 0.9692307692307692

score for the test set 0.6585365853658537

Accuracy of the model 0.6585365853658537

Graphical Representation of Confusion matrix :



4.4 SVM :

Support Vector Machine (SVM) is a powerful algorithm used in machine learning for classification and regression tasks. The goal of SVM is to find the best possible boundary that separates the data into different classes. This boundary is called a hyperplane, which can be linear or non-linear, depending on the type of SVM used. The SVM algorithm works by mapping the data points into a high-dimensional feature space and then finding the hyperplane that maximizes the distance between the closest points from each class. These closest points are called support vectors, and the distance between them is called the margin.

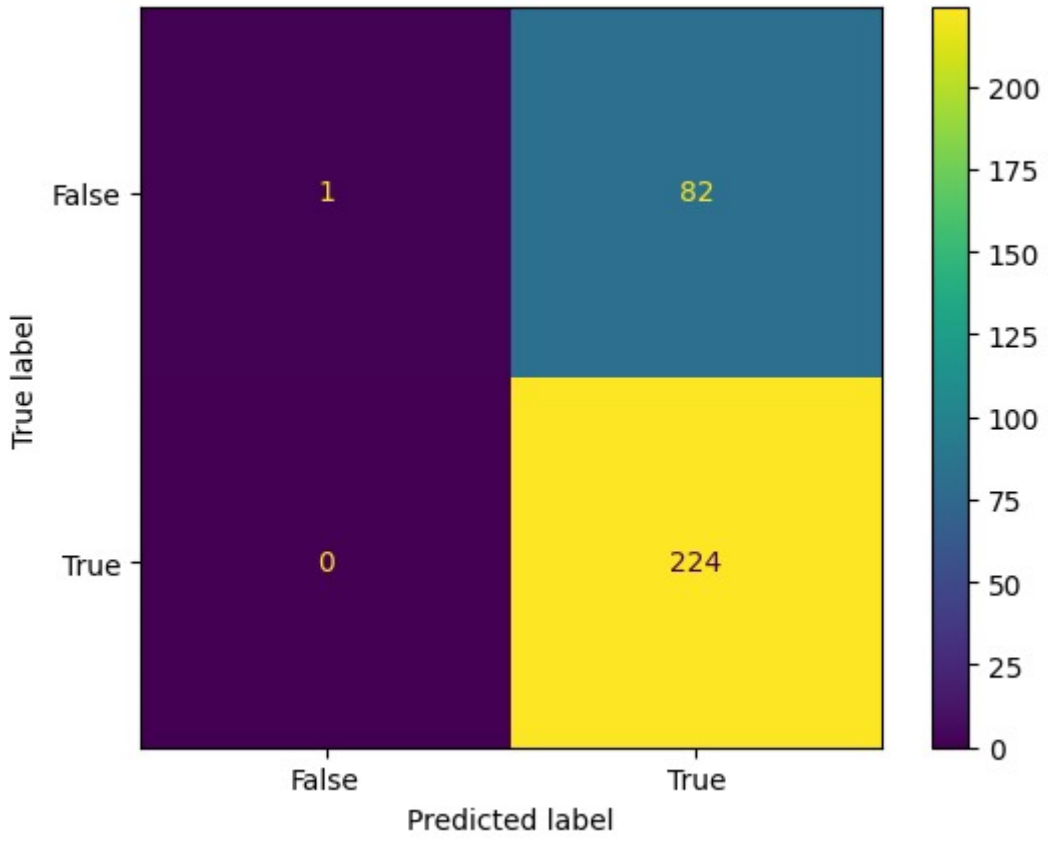
Accuracy for SVM:

accuracy=accuracy_score(yp,y_test)

print(accuracy)

0.7328990228013029

Graphical Representation of Confusion matrix :



4.5 Naïve-Bayes :

Bayes Theorem is a Bayes' Theorem is a fundamental concept in probability theory, named after the Reverend Thomas Bayes. It provides a way to update our beliefs about the likelihood of an event occurring, based on new evidence or information that we receive.Bayes' Theorem is a fundamental concept in probability theory, named after the Reverend Thomas Bayes. It provides a way to update our beliefs about the likelihood of an event occurring, based on new evidence or information that we receive.Bayes' Theorem states that the probability of a hypothesis H, given some observed evidence E, is proportional to the prior p Bayes' Theorem is a fundamental concept in probability theory, named after the Reverend Thomas Bayes. It provides a way to update our beliefs about the likelihood of an event occurring, based on new evidence or information.

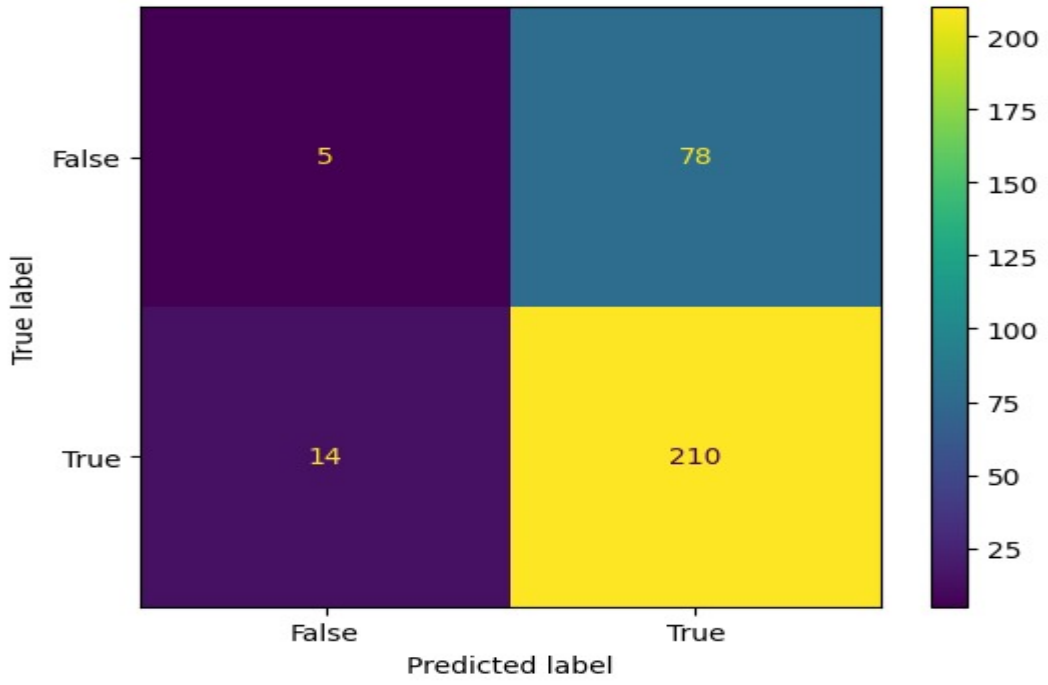
Accuracy for bayes:

accuracy=accuracy_score(yp,y_test)

print(accuracy)

0.7003257328990228

Graphical Representation of Confusion matrix :



5 . Results and Discussion

Accuracy through Logistic Regression:

Ml model	Accuracy
Logistic Regression	0.7328990228013029

Accuracy through Decision Tree:

Ml model	Accuracy
Decision tree	0.8371335504885994

Accuracy through KNN:

Ml model	Accuracy
KNN	0.6585365853658537

Accuracy through Navies Bayes Theorem:

Ml model	Accuracy
Navies Bayes theorem	0.7003257328990228

Accuracy through SVM:

Ml model	Accuracy
Support vector machine	0.7328990228013029

6.CONCLUSION AND FUTURE SCOPE

This application is working properly and meeting to all Banker requirements. This component can be easily plugged in many other systems. It works correctly and fulfills all requirements of bankers and can be connected to many other systems. There were multiple malfunctions in the computers, content errors and fixing of weight in computerized prediction systems.In the near term, the banking software could be more reliable, accurate, and dynamic in nature and can be fit in with an automated processing unit. There have been numbers cases of computer glitches, errors in content and most important weight of features is fixed in automated prediction system more secure, reliable and dynamic weight adjustment. The system is trained on old training dataset in future software can be made such that new testing date should also take part in training data after some fix time. Machine learning helps to understand the factors which affect the specific outcomes most. Other models like neutral network and discriminate analysis can be used individually or combined for enhancing reliability and accuracy prediction.

This project helped us to learn about the complicated system of the loan prediction system and the best model that can work with this particular project. It works correctly and fulfills all requirements of bankers. This system properly and accurately calculates the result. It predicts the loan is approve or reject to loan applicant or customer very accurately.

7.REFERENCES

[1] Kumar Arun, Garg Ishan, Kaur Sanmeet ,“Discuss function of ML in banking system”, Loan Approval Prediction based on Machine Learning Approach, Volume 18, Issue 3, Ver. I,e-ISSN: 2278-0661,p-ISSN: 2278-8727,MayJun. 2016.

[2] Shiva Agarwal, “Describe the concepts of data mining”, Data Mining: Data Mining Concepts and Techniques ,INSPEC Accession Number: 14651878,Electronic ISBN:978- 0-7695-5013-8, 2013. [3] Aboobyda, J. H., and M. A. Tarig. "Developing Prediction Modelof Loan Risk in Banks Using Data Mining." Machine Learning and Applications: An International Journal (MLAIJ)3.1, 2016.

[4] A kindaini, Bolarinwa. “Machine learning applications in mortgage default prediction.”University of Tampere, 2017.

[5] Amir E. Khandani, Adlar J. Kim and Andrew Lo, “Consumer credit-risk modelsvia machinelearning algorithms and risk management in banking system”,J. Bank Financ., vol.34, no. 11,pp. 27672787, Nov. 2010.

[6] Aurélien Géron, “Focus on implementing ML programs using the library Scikit-Learn”,Publisher – O‟Reilly Media, Inc, Edition – Second Edition.

[7] Yuxi (Hayden) Liu, “discuss about the Python Programming”, Python Machine Learning By Example, Edition – Third Edition, Publisher – Packt Publishing.

[8] Jiawei Han, Micheline Kamber, Jian Pei , “study the concepts of data mining”, Data Mining Concepts and Techniques , Edition- third edition,

[9] Ted Dunning, Ellen Friedman, “discuss logistic regression principles”,

Machine Learning Logistics, ISBN: 9781491997611Publisher(s): O'Reilly Media, Inc, Released October 2017.

[10]Lior Rokach , “discuss Data mining and decision tree” ,Data Mining with Decision Trees: Theory and Applications: Theory and Applications 81 (Series In Machine Perception And Artificial Intelligence) , Second Edition ,23 October 2014.