CHAPTER 1

INTRODUCTION

Individuals working in banks face a bunch of issues which include endorsement of a credit. In the 21st century, individuals regularly depend on innovation to handle such issues. This paper proposes an advance authorizing framework which decides if an advance ought to be given to an individual, in view of specific characteristics. Regardless of banks adhering to tough guidelines and guidelines and leading fastidious individual verifications while authorizing a credit and remembering the likelihood of the individual's capacity to restore the advance, frequently such circumstances are confronted where in, the individual can't reimburse the advance that has been given to him. In this paper, the framework that we propose for the financiers will assist them with anticipating the solid clients who have applied for advance, in this manner improving the odds of their advances being reimbursed in time. This arrangement is finished utilizing Guileless Bayesian calculation. So as to improve the grouping exactness, the nature of the information is improved before ordering it by utilizing K-NN and Binning calculations. This framework utilizes these calculations so as to yield a superior effectiveness in order to lessen the chance of such an issue. The proposed framework moreover encourages self-affirmation with respect to the equivalent for the average person.

The primary issue that we attempt to explain in this undertaking is to foresee the credit default rate. Precise forecast of whether an individual will default on their advance, and how much misfortune it will cause has a reasonable significance for banks' hazard the board. These days, banks have remembered a lot of data for its assessment of credit issuance, and a portion of these data has an obscure causal relationship with the advance default rate. The developing measure of information because of improved information catch and information stockpiling innovation has welcome us to another point of view on this difficult which is utilized to be practiced by budgetary and monetary investigation.

Subsequently, in this venture, we would apply our insight from information mining class and test an assortment of information mining approaches on this issue. This information mining task, in nature, is a relapse task as the objective quality, advance default rate is a consistent numerical worth. Be that as it may, as we improve our model, there are a few varieties.

This Issue is finished by mining the Large Information of the past records of the individuals to whom the advance was conceded previously and based on these records/encounters the machine was prepared utilizing the AI model which give the most exact outcome. The principle target of this paper is to anticipate in the case of relegating the credit to a specific individual will be sheltered or not. We have executed this advance expectation issue utilizing Choice tree calculation and information cleaning in Python as there are missing qualities in the dataset. We use map work for the missing qualities. The point of this paper is to apply AI procedure on dataset which has 1000 cases and 7 numerical and 6 clear cut characteristics. The noteworthiness of a client for endorsing advance rely upon a few parameters, for example, record of loan repayment, portion and so forth.

Because of gigantic development in information the financial business manages, examination and change of the information into helpful information has become an undertaking past human capacity. Information mining methods can be received in taking care of business issues by discovering examples, affiliations and connections which are covered up in the business data put away in the information bases. By utilizing information mining methods to break down examples and patterns, bank officials can anticipate, with expanded precision, how clients will respond to changes in financing costs, which clients are probably going to acknowledge new item offers, which clients will be at a higher hazard for defaulting on a credit, and how to make client connections increasingly gainful. Globalization and the firm rivalry had driven the banks center towards client maintenance and extortion avoidance. To help them for the equivalent, information mining is utilized. By dissecting the past information, information mining can help banks to anticipate solid clients. In this manner they can forestall fakes, they can likewise get ready for propelling diverse exceptional proposals to hold those clients who are valid. Certain regions that successfully use information mining in banking industry are showcasing, chance administration and client relationship the board.

Risk Management: It is generally utilized for overseeing dangers in the financial business. Bank administrators need to know the believability of clients they are managing. Offering new clients charge cards, broadening existing clients' credit extensions, and favoring advances can be hazardous choices for banks, on the off chance that they know nothing about their clients. Banks give advances to their clients by checking the different subtleties identifying with the credit, for example, measure of advance, loaning rate, reimbursement period and so forth.

Despite the fact that, banks are careful while giving advance, there are odds of advance reimbursing defaults by clients. Information mining method assists with recognizing borrowers who reimburse advances immediately from the individuals who default.

Customer Relationship Management: Information mining can be helpful in all the three periods of a client relationship cycle, for example, client procurement, expanding estimation of the client and client maintenance. Client securing and maintenance are significant worries of any industry, particularly the financial business. Banks need to cook the necessities of the clients by offering the types of assistance they like. This will at last lead to client reliability and client maintenance. Information mining methods help to break down the clients who are faithful from the individuals who move to different banks for better administrations. On the off chance that the client is moving from his bank to another, explanations behind such moving and the last exchange performed before moving can be known, and this will assist the manages an account with performing better and hold their clients.

1.1 PURPOSE OF STUDY:

In this day and age, getting advance from money related organizations has become a typical wonder. Consistently numerous individuals apply for credits, for an assortment of direction. Be that as it may, not all the applications are solid, and not every person can be endorsed. Consistently, there are situations where individuals don't reimburse the heft of the advance add up to the bank which brings about gigantic money related misfortune. The hazard related with settling on a choice on a credit endorsement is tremendous. Thus, the possibility of this undertaking is to assemble advance information and utilizing AI procedures on this information to remove significant data and foresee if a client would have the option to reimburse the credit or not. As it were, the objective is to anticipate if the client would be a defaulter or not.

1.2 PROBLEM STATEMENT:

Multiclass classification problem for predicting loan status from rich dataset. The problem can be reduced to binary classification problem to build a loan approval pre-check system for potential customers.

1.3 MOTIVATION:

The advance is one of the most significant results of the banking. All the banks are attempting to make sense of viable business techniques to convince clients to apply their advances. Be that as it may, there are a few clients act adversely after their application are affirmed. To forestall this circumstance, banks need to discover a few techniques to foresee clients' practices. AI calculations have a really decent exhibition on this reason, which are broadly utilized by the banking. Here, I will chip away at advance practices forecast utilizing AI models.

1.4 METHODOLOGY:

In this project, different common language preparing strategies and AI calculation to group credit information utilizing the library from python are utilized and is actualized utilizing python programming language. Perusing the train, test and approval information records and playing out some pre-handling. Preparing information will be pre-handled and includes are closed. These pre-handled preparing information and highlights are exposed to AI calculation, from which the best appropriate calculation is utilized to fabricate the model like choice tree, strategic relapse. Last characterization model is manufacture and likelihood of bogus and truth can be identified in the yield.

CHAPTER 2

SYSTEM REQUIREMENTS AND LANGUAGE USED

2.1 HARDWARE AND SOFTWARE REQUIREMENTS

Hardware System Configuration:

Processor	- Intel Core i5
Speed	- 1.8 GHz
RAM	- 256 MB (min)
Hard Disk	- 10 GB

Software System Configuration:

Operating System	- windows 10
Programming Language	- machine learning with python
Compiler	- anaconda

2.2 ABOUT THE LANGUAGE

The language utilized here is python with different normal language preprocessing strategies and AI calculation and bundles. Python is the significant level, deciphered, broadly useful programming language. Made by Guido van Rossum, discharged in 1991. Python plan theory stresses code meaningfulness with its eminent utilization of noteworthy whitespace. It language builds and item arranged methodology plan to assist developers with composing clear consistent code for little and huge scope ventures. This language additionally bolsters different programming ideal models, including object situated, and work programming. Python can be utilized for creating work area GUI applications, sites and web application. Python is utilized as means for making apparatuses and models, rationale framework in games, just as for bringing in or sending out records and mechanizing undertakings. GIMP is a raster illustrations manager that mostly underpins working with vectors. It is likewise used to compose extra modules, for instance, channels. This language is constantly portrayed as a "batteries included" language because of its exhaustive standard library.

As AI calculation is utilized in this undertaking, python is a generally mainstream, universally useful programming language appropriate for an assortment of assignments in AI. Python consolidates noteworthy force with extremely clear linguistic structure. Python is likewise usable as an expansion language for applications written in other language that need simple to-utilize scripting or robotization interfaces. Python is generally considered as the favored language for instructing and learning AI.

CHAPTER 3

SYSTEM DESIGN

3.1 ARCHITECTURE

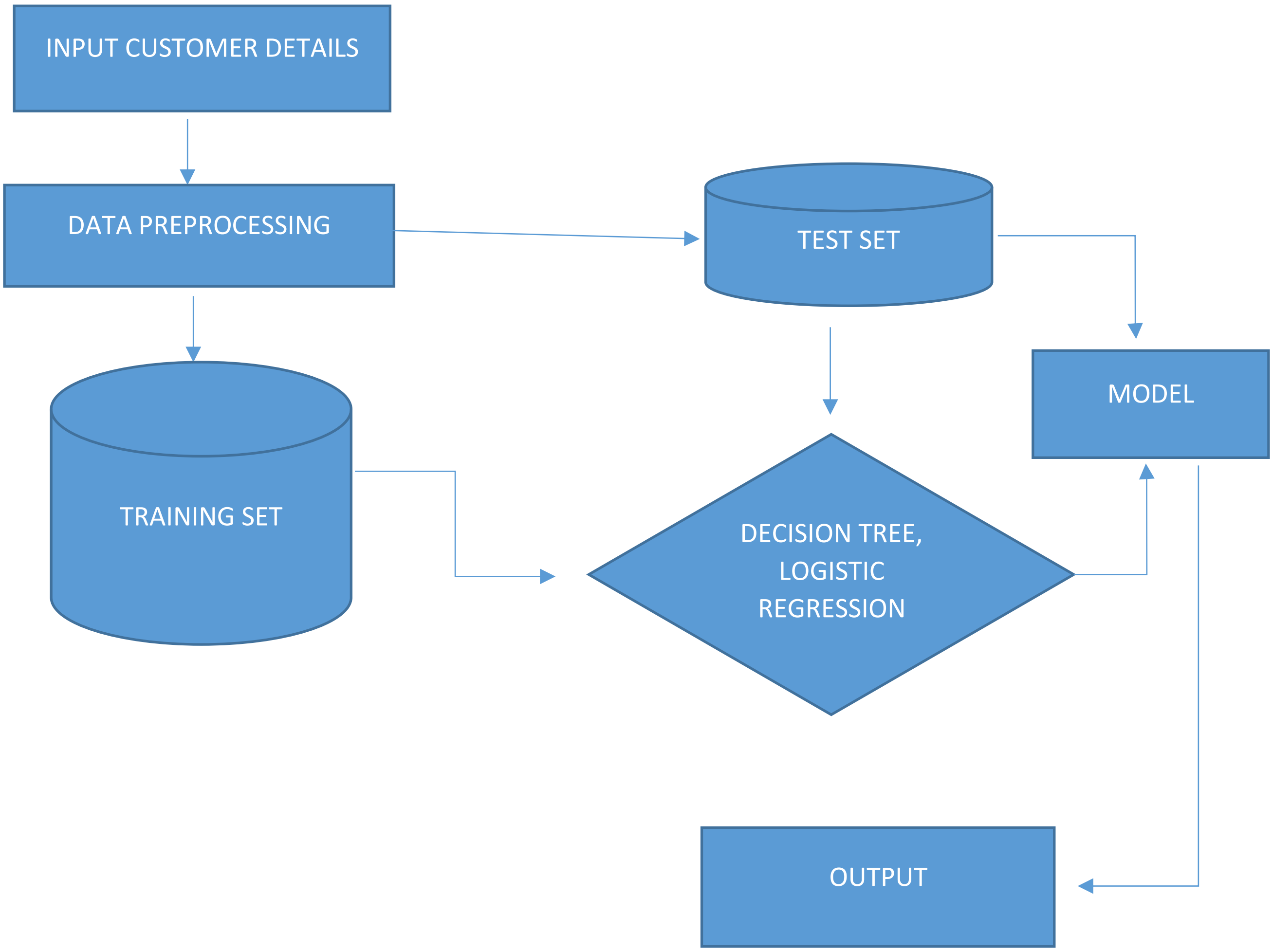


Fig 3.1: ARCHITECTURE OF LOAN PREDICTION

1. INPUT: The principle feature of this credit validity forecast framework is that it utilizes choice tree, calculated relapse enlistment information mining calculation to sift through the advance solicitations. A choice tree is created by preforming information mining on a current bank dataset containing 4520 records and 17 properties.

2. DATA PREPROCESSING: At first the ascribes which are basic to make a credit believability forecast is related to data gain as the property evaluator and ranker as the hunt technique. Manual pre-handling is likewise performed.
3. DATA FITERING: Last dataset after pre-handling is partitioned so that there is 66% preparing set and 34% test set. Test set is utilized to approve the conclusive outcome of the performed.
4. DECISION TREE ALGORITHM: A productive choice tree is detailed with choice tree enlistment calculation. It delivers a model with the most pertinent 6 property. Characteristic with rank-1 is put as the root hub of the choice tree, different credits from rank-2 to rank-6 comprise the moderate hubs. A choice is made at every hub and the leaf hub gives us the conclusive outcome. That is, on the off chance that the client have the base advance reimbursement limit, at that point the future dangers can be maintained a strategic distance from. The principle advantage of applying information mining is that we can generally depend on the aftereffect of the calculation to acknowledge or dismiss the advance application.
5. LOGISTIC REGRESSION: This is a measurable model that in its essential structure utilizes a strategic capacity to show a double reliant variable, albeit a lot increasingly complex expansion exists. In relapse examination, strategic relapse is assessing the parameters of a calculated model.

3.2 ALGORITHM

- Step1 : start
- Step2 : recording the loan data(ID , married , loan amount , education Etc.)
- Step3 : collecting the dataset from the loan recorded.
- Step4 : the data is trained.
- Step5 : decision tree(taken to predict the target node),then from the target node we are predicting the loan approval based on the logistic regression (this is a basic form uses a logistic function to model a binary dependent variable).
- Step6 : the data is tested

3.3 FLOW CHART

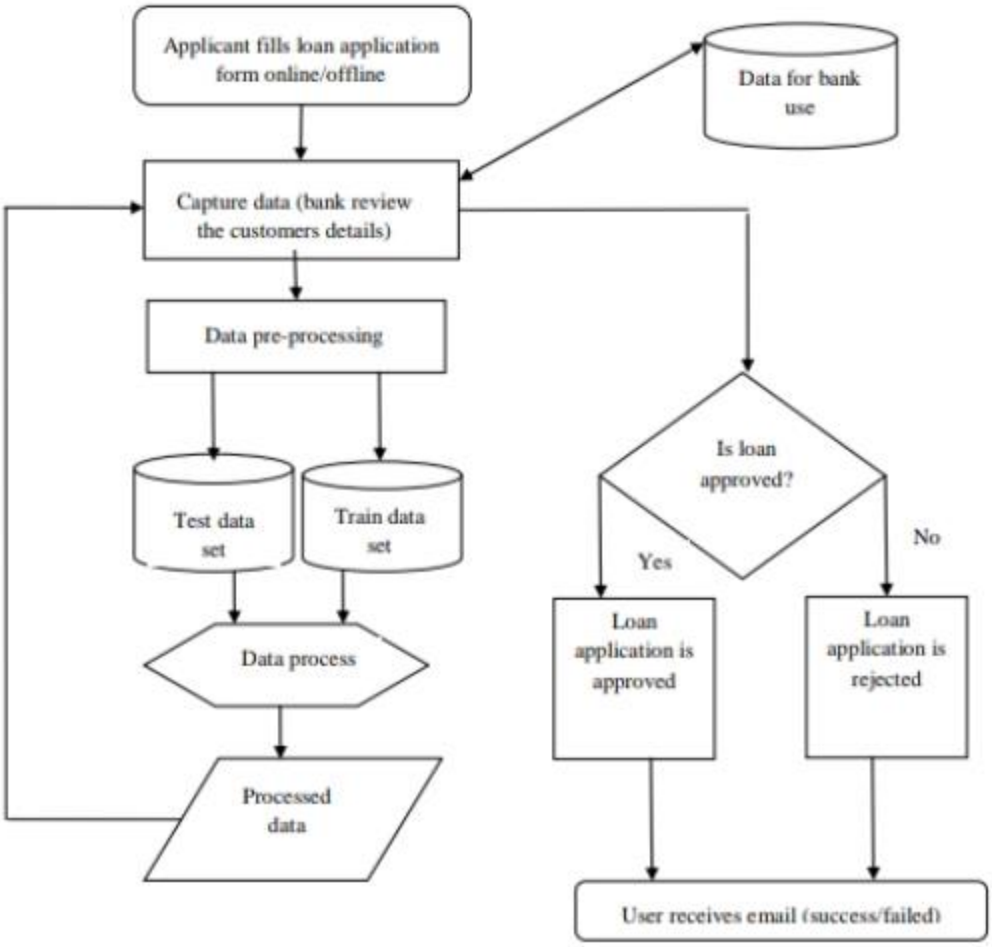


Fig 3.3: FLOW CHART OF LOAN PREDICTION

3.4 CODE

```
import pandas as pd

import numpy as np

Train=pd.read_csv(r'C:\Users\PREETHI\Desktop\Train.csv')

Train.Loan_Status=Train.Loan_Status.map({'Y':1,'N':0})

Train.isnull().sum()

Loan_status=Train.Loan_Status

Train.drop('Loan_Status',axis=1,inplace=True)

test=pd.read_csv(r'C:\Users\PREETHI\Desktop\test.csv')

Loan_ID=test.Loan_ID

data=Train.append(test)

data.head()

data.describe()

data.isnull().sum()

data.Dependents.dtypes

import matplotlib.pyplot as plt

import seaborn as sns

get_ipython().magic('matplotlib inline')

corrmat=data.corr()

f,ax=plt.subplots(figsize=(9,9))

sns.heatmap(corrmat,vmax=.8,square=True)

data.Gender=data.Gender.map({'Male':1,'Female':0})

data.Gender.value_counts()
```

```
corrmat=data.corr()

f,ax=plt.subplots(figsize=(9,9))

sns.heatmap(corrmat,vmax=.8,square=True)

data.Married=data.Married.map({'Yes':1,'No':0})

data.Married.value_counts()

data.Dependents=data.Dependents.map({'0':0,'1':1,'2':2,'3+':3})

data.Dependents.value_counts()

corrmat=data.corr()

f,ax=plt.subplots(figsize=(9,9))

sns.heatmap(corrmat,vmax=.8,square=True)

data.Education=data.Education.map({'Graduate':1,'Not Graduate':0})

data.Education.value_counts()

data.Self_Employed=data.Self_Employed.map({'Yes':1,'No':0})

data.Self_Employed.value_counts()

data.Property_Area.value_counts()

data.Property_Area=data.Property_Area.map({'Urban':2,'Rural':0,'Semiurban':1})

data.Property_Area.value_counts()

corrmat=data.corr()

f,ax=plt.subplots(figsize=(9,9))

sns.heatmap(corrmat,vmax=.8,square=True)

data.head()

data.Credit_History.size

data.Credit_History.fillna(np.random.randint(0,2),inplace=True)
```

```
data.isnull().sum()

data.Married.fillna(np.random.randint(0,2),inplace=True)

data.isnull().sum()

data.LoanAmount.fillna(data.LoanAmount.median(),inplace=True)

data.Loan_Amount_Term.fillna(data.Loan_Amount_Term.mean(),inplace=True)

data.isnull().sum()

data.Gender.value_counts()

from random import randint

data.Gender.fillna(np.random.randint(0,2),inplace=True)

data.Gender.value_counts()

data.Dependents.fillna(data.Dependents.median(),inplace=True)

data.isnull().sum()

corrmat=data.corr()

f,ax=plt.subplots(figsize=(9,9))

sns.heatmap(corrmat,vmax=.8,square=True)

data.Self_Employed.fillna(np.random.randint(0,2),inplace=True)

data.isnull().sum()

data.head()

data.drop('Loan_ID',inplace=True,axis=1)

data.isnull().sum()

train_X=data.iloc[:614,]

train_y=Loan_status

X_test=data.iloc[614:,]
```

```
seed=7

from sklearn.model_selection import train_test_split

train_X,test_X,train_y,test_y=train_test_split(train_X,train_y,random_state=seed)

from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

from sklearn.linear_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.svm import SVC

from sklearn.neighbors import KNeighborsClassifier

from sklearn.naive_bayes import GaussianNB

models=[]

models.append(("logreg",LogisticRegression()))

models.append(("tree",DecisionTreeClassifier()))

models.append(("lda",LinearDiscriminantAnalysis()))

models.append(("svc",SVC()))

models.append(("knn",KNeighborsClassifier()))

models.append(("nb",GaussianNB()))

seed=7

scoring='accuracy'

from sklearn.model_selection import KFold

from sklearn.model_selection import cross_val_score

result=[]

names=[]

for name,model in models:
```

```
#print(model)
kfold=KFold(n_splits=10,random_state=seed)
cv_result=cross_val_score(model,train_X,train_y,cv=kfold,scoring=scoring)
result.append(cv_result)
names.append(name)
print("%s %f %f" % (name,cv_result.mean(),cv_result.std()))
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
svc=LogisticRegression()
svc.fit(train_X,train_y)
pred=svc.predict(test_X)
print(accuracy_score(test_y,pred))
print(confusion_matrix(test_y,pred))
print(classification_report(test_y,pred))
df_output=pd.DataFrame()
outp=svc.predict(X_test).astype(int)
outp
df_output['Loan_ID']=Loan_ID
df_output['Loan_Status']=outp
df_output.head()
import pandas as pd
import numpy as np
```

```
loan = pd.read_csv(r'C:\Users\SWATHI\Desktop\train.csv',encoding='utf-8')

loan.Gender = loan.Gender.fillna('Male')

loan.Married=loan.Married.fillna('Yes')

loan.Dependents=loan.Dependents.fillna('0')

loan.Self_Employed=loan.Self_Employed.fillna('No')

loan.LoanAmount=loan.LoanAmount.fillna(loan.LoanAmount.mean())

loan.Loan_Amount_Term=loan.Loan_Amount_Term.fillna('360.0')

loan.Credit_History=loan.Credit_History.fillna('1.0')

loan.Education=loan.Education.replace({'Graduate':1,'Not Graduate':0})

loan.Self_Employed=loan.Self_Employed.replace({'Yes':1,'No':0})

loan.Dependents=loan.Dependents.replace({'3+':3})

loan.Property_Area=loan.Property_Area.replace({'Urban':2,'Rural':0,'Semiurban':1})

loan.Loan_Status=loan.Loan_Status.replace({'Y':1,'N':0})


loan = loan.drop(['Loan_ID'],axis=1)

temp = loan

temp.ApplicantIncome = pd.DataFrame(temp.ApplicantIncome)

y = -1

for x in temp.ApplicantIncome:

    y = y+1

    if (x > 0 and x < 5000):

        temp.ApplicantIncome[y] = 0

y = -1
```

```
for x in temp.ApplicantIncome:
    y = y+1
    if (x >= 5000 and x < 10000):
        temp.ApplicantIncome[y] = 1
y = -1
for x in temp.ApplicantIncome:
    y = y+1
    if  x >= 10000:
        temp.ApplicantIncome[y] = 2
y = -1
for x in temp.CoapplicantIncome:
    y = y + 1
    if x > 0:
        temp.CoapplicantIncome[y] = 1
temp.Loan_Amount_Term = temp.Loan_Amount_Term.astype(float)
loan.Gender=loan.Gender.replace({'Male':1,'Female':0})
loan.Married=loan.Married.replace({'Yes':1,'No':0})
from sklearn.model_selection import train_test_split
# y = f(x)
X = temp[temp.columns[:-1]] # all but last column --- contains feature columns
Y = temp[temp.columns[-1]]  # last column  --- contains target column
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.33, random_state=50)
X_train.shape
```

```
Y_train.shape

X_test.shape

Y_test.shape

from sklearn.linear_model import LogisticRegression

loan_model = LogisticRegression()

# Fitting Logistic Regression to the Training set

loan_model.fit(X_train, Y_train)

Y_pred = loan_model.predict(X_test)

z = loan_model.score(X_test, Y_test)

import pickle

saved_model = pickle.dumps(loan_model)

loan_model_from_pickle = pickle.loads(saved_model)

print(z)

from sklearn.externals import joblib

from joblib import dump

dump(loan_model, 'file.joblib')

from joblib import load

new_loanapp = load('file.joblib')

new_loan_application = [0, 1, 1, 0, 1, 0, 10.0, 700.0, 360.0, 1, 1]

p = new_loanapp.predict([new_loan_application])

print(p)
```

CHAPTER 4

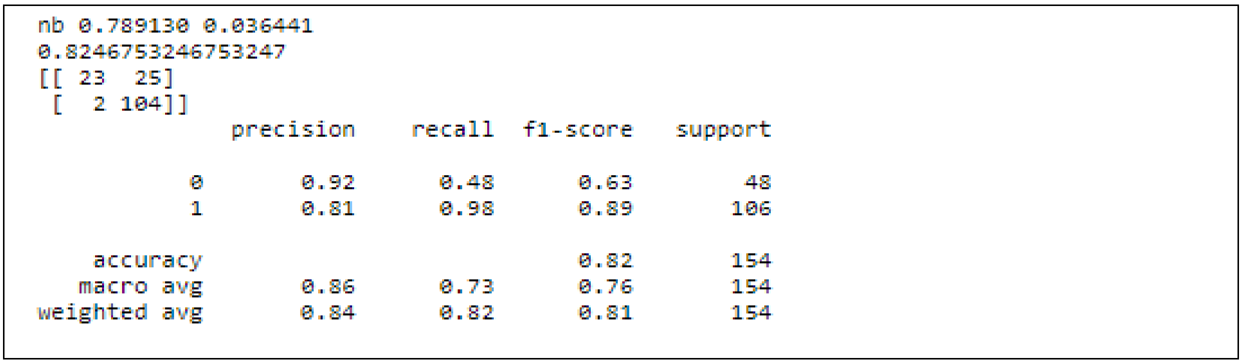
RESULTS AND DISCUSSION

4.1 SUMMARY OF RESSULTS OBTAINED

In this project we are dividing the given data into train and test data. Using machine learning algorithm like decision tree and logistic regression we are training the data. Based on the train data, the test data is tested, were this predicts the loan status for all the given dataset

Based on the accuracy score the model is build in such a way that it predicts for the individual data. The output is in the form of 0 and 1(binary form), were '1' indicates that loan is approved and '0' for loan rejected.

4.2 OUTPUT



```
nb 0.789130 0.036441
0.8246753246753247
[[ 23  25]
 [  2 104]]
              precision    recall  f1-score   support

           0       0.92      0.48      0.63        48
           1       0.81      0.98      0.89       106

    accuracy                           0.82       154
   macro avg       0.86      0.73      0.76       154
weighted avg       0.84      0.82      0.81       154
```

Fig 4.2.1: ACCURACY SCORE

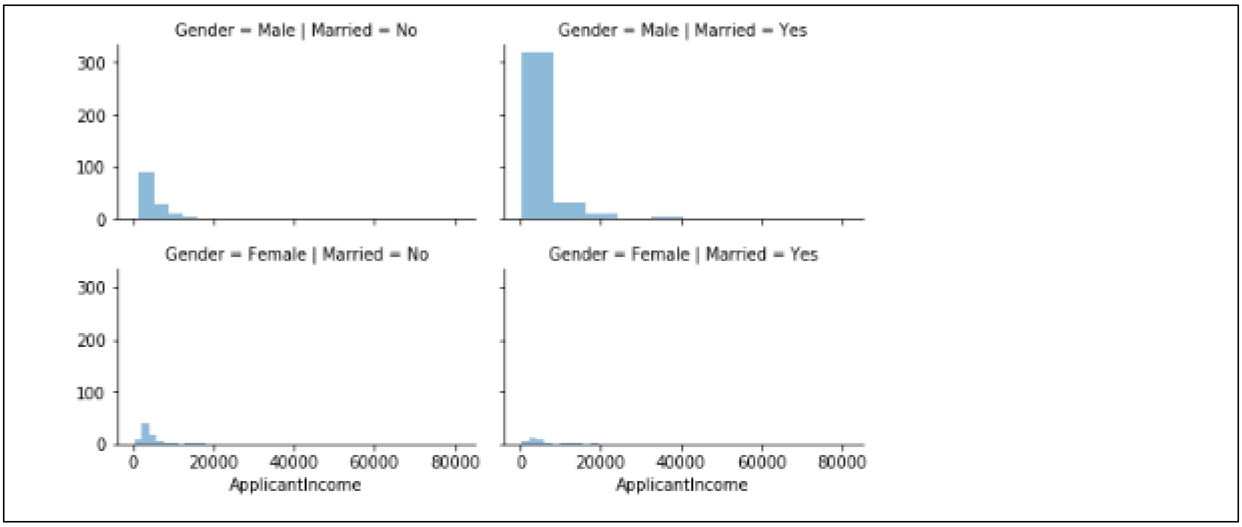


Fig 4.2.2: GRAPH BETWEEEN GENDER AND MARRIED

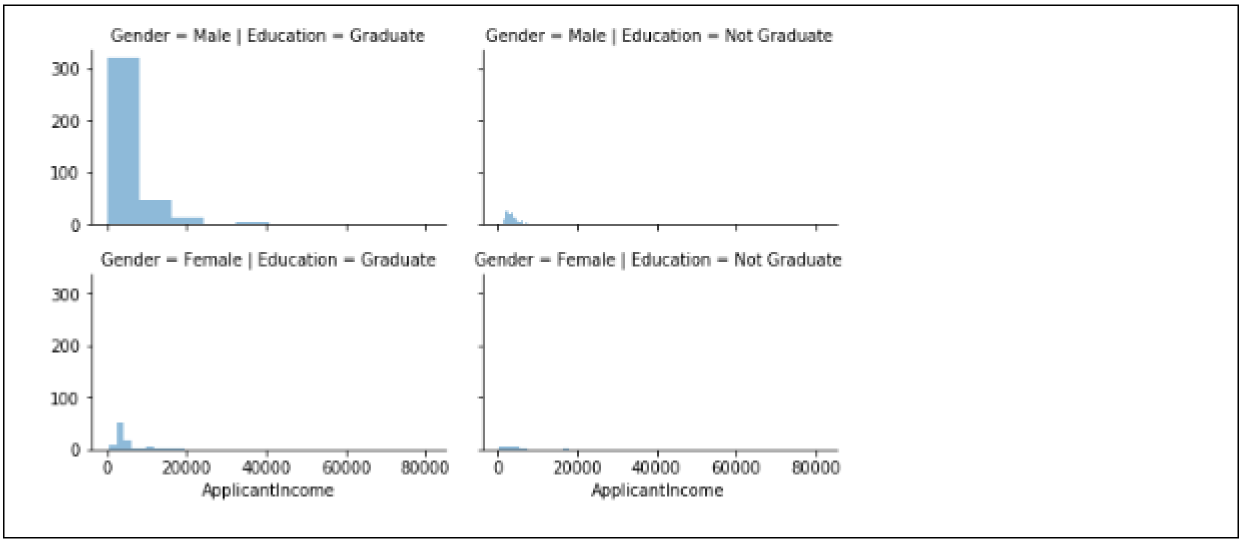


Fig 4.2.3: GRAPH BETWEEN GENDER AND EDUCATION

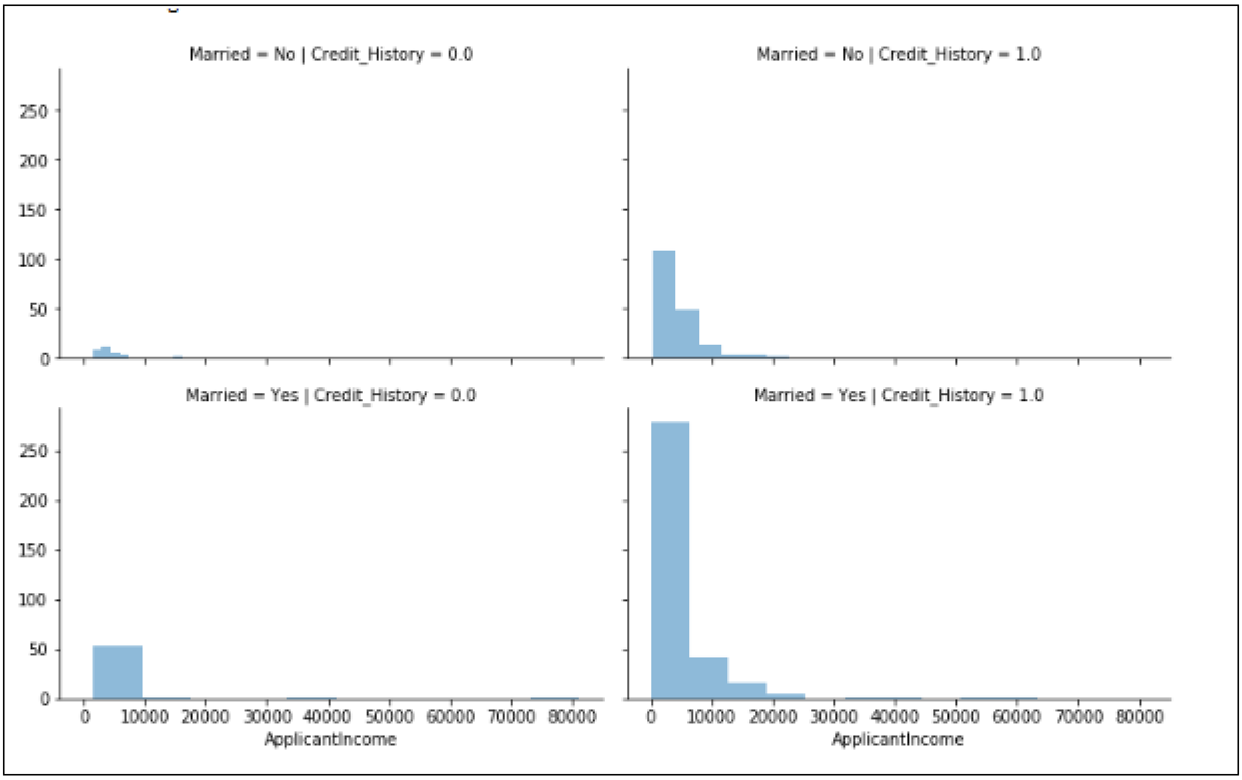


Fig 4.2.4: GRAPH BETWEEN MARRIED AND CREDIT_HISTORY

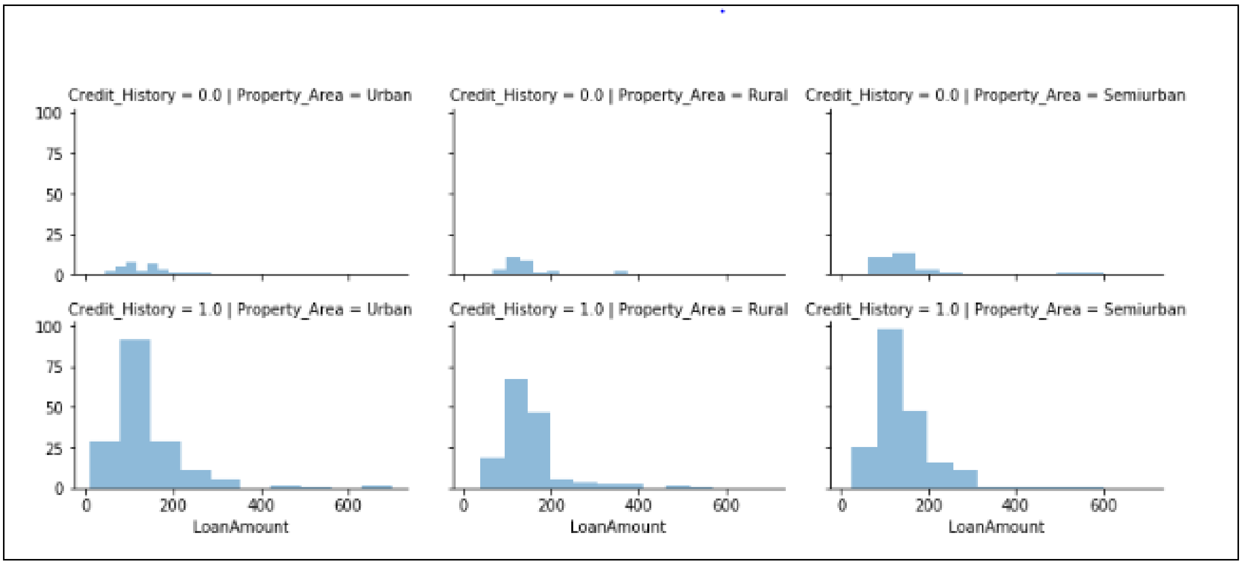


Fig 4.2.5: GRAPH BETWEEN CREDIT_HISTORY AND PROPERTY_AREA

	Loan_ID	Loan_Status
0	LP001015	1
1	LP001022	1
2	LP001031	1
3	LP001035	1
4	LP001051	1

Fig 4.2.6: LOAN STATUS FOR ALL THE DATASET

```
0.7881773399014779
[1]
```

Fig4.2.7: LOAN STATUS FOR AN INDIVIDUAL DATA

CHAPTER 5

CONCLUSION

This application can help banks in anticipating the fate of credit and its status and relies upon that they can make a move in introductory long periods of advance. Utilizing this application banks can diminish the quantity of awful advances from bringing about cut off misfortunes. A few AI calculations and bundles were utilized to set up the information and to fabricate the arrangement model. AI bundle libraries help in fruitful information examination and highlight determination. Utilizing this technique bank can without much of a stretch distinguish the necessary data from immense measure of informational collections and aides in fruitful advance forecast to diminish the quantity of awful credit issues. Information mining strategies are helpful to the financial part for better focusing on and procuring new clients, most significant client maintenance, programmed credit endorsement which is utilized for extortion avoidance, misrepresentation identification progressively, giving section based item, investigation of the client, exchange designs after some time for better maintenance and relationship, hazard the executives and showcasing.

REFERENCES

www.ijetjournal.org/volume5/issue2/IJET-V5I2P28.pdf

www.slideshare.net/AliCloud/loan-default...

www.datasciencecentral.com/profiles/blogs/loan...

www.youtube.com/watch?v=TJOhBlp4fKA

www.coursera.org/lecture/ml-classification/...