A PROJECT ETE REPORT

on

LAON APPROVAL PREDICTION USING MACHINE LEARNING

Submitted in partial fulfillment of the

requirement for the award of the

degreeof

Bachelors Of Technology (Computer Science And Engineering)

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of Mr.S Prakash

Submitted By

SANJAY KUMAR

KHARBIND(19SCSE1010821)

GANESH YADAV(19SCSE1010806)

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

DECEMBER,
2021

Abstract

Loan approval is a very important process for banking organizations. The system approved or reject the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. In recent years many researchers worked on loan approval prediction systems. Machine Learning (ML)techniques are very useful in predicting outcomes for large amount of data. In this paper three machine learning algorithms, Logistic Regression(LR), Decision Tree (DT) and Random Forest (RF)are applied to predict the loan approval of customers. The experimental results conclude that the accuracy of Decision Tree machine learning algorithm is better as compared to Logistic Regression and Random Forest machine learning approaches. In our banking system, bank have many products to sell but main source of income is its credit line. So, they can earn from interest of those loans which they credit. Loan approval is very important process is for banking organization as well as applicants. But the recovery of loan is very important things for bank because if the bank is not able to recover its amount, then bank are automatically goes to loss. Nowadays Fraud system is increasing day – by -day so to control fraud system in our bank we have to implements some Machine learning algorithm to check documentation of an applicant's online by filling form. After checking all the document of applicant/client, if they satisfy all the requirement of form then he/she can get loan approval otherwise they reject. Machine Learning (ML)technique are very useful in predicting outcome for large amount of data. In this paper some machine learning algorithms with some python library like Pandas, NumPy, Matplotlib, seaborn and some Machine Learning like support Vector Machine and Naïve Bayes are apply to predict the loan approval of customers. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters (i) collection of data (ii) Data cleaning and (iii) performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting.

TABLE OF CONTENTS

CHAPTER-1

Introduction

1.1 Formulation of the Problem: -

In our banking system, bank have many products to sell but main source of income is its credit line. So, they can earn from interest of those loans which they credit. Loan approval is very important process is for banking organization as well as applicants. But the recovery of loan is very important things for bank because if the bank is not able to recover its amount, then bank are automatically goes to loss. Nowadays Fraud system is increasing day – by -day so to control fraud system in our bank we have to implements some Machine learning algorithm to check documentation of an applicant's online by filling form. After checking all the document of applicant/client, if they satisfy all the requirement of form then he/she can get loan approval otherwise they reject. Machine Learning (ML)technique are very useful in predicting outcome for large amount of data.

Now a day's people rely on bank loans to fulfill their needs. The rate of loan applications increases with a very fast speed in recent years. Risk is always involved in approval of loans. The banking officials are very conscious about the payment of the loan amount by its customers. Event after taking lot of precautions and analyzing the loan applicant data, the loan approval decisions are not always correct. There is need of automation of this process so that loan approval is less risky and incur less loss for banks

Artificial Intelligence AI is an emerging technology now a day. The application of AI solves many problems of the real world. Machine Learning is an AI technique which is very useful in prediction systems. Figure 1 is showing a basic model of machine learning. It creates a model from a training data. While making the prediction the model which is developed by training algorithm (which is machine learning) is used. The machine learning algorithm trained the system using a fraction of the data available and test the remaining data.

The machine Learning techniques can be applied on a sample test data first and then can be used in making prediction related decisions. This paper applied the machine learning approaches in solving loan approval problem of banking sector.

1.2 Tools and Technology Used: -

Data:- In computing, data is information that has been translated into a form that is efficient for movement or processing. Relative to today's computers and transmission media, data is information converted into binary digital form. It is acceptable for data to be used as a singular subject or a plural subject. Raw data is a term used to describe data in its most basic digital format**.**

Random Forecast:- Random Forest (RF) is a very useful machine learning algorithm. It is mostly used in areas such as classification, regression analysis etc. At the training time RF algorithm creates many decision trees.

RF is a supervised learning approach which need a test data for the model for training. It creates random forests for the problem set and then find the solution using these random forests.

Decision Tree Algorithm: - The decision tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for regression problem. The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree. Decision TREE is a supervised ML technique which is non parametric in nature. It has predefined target variable which is generally used in problem classification. It is useful for classification and regression both. It works categorical & continuous both for input and output variables.

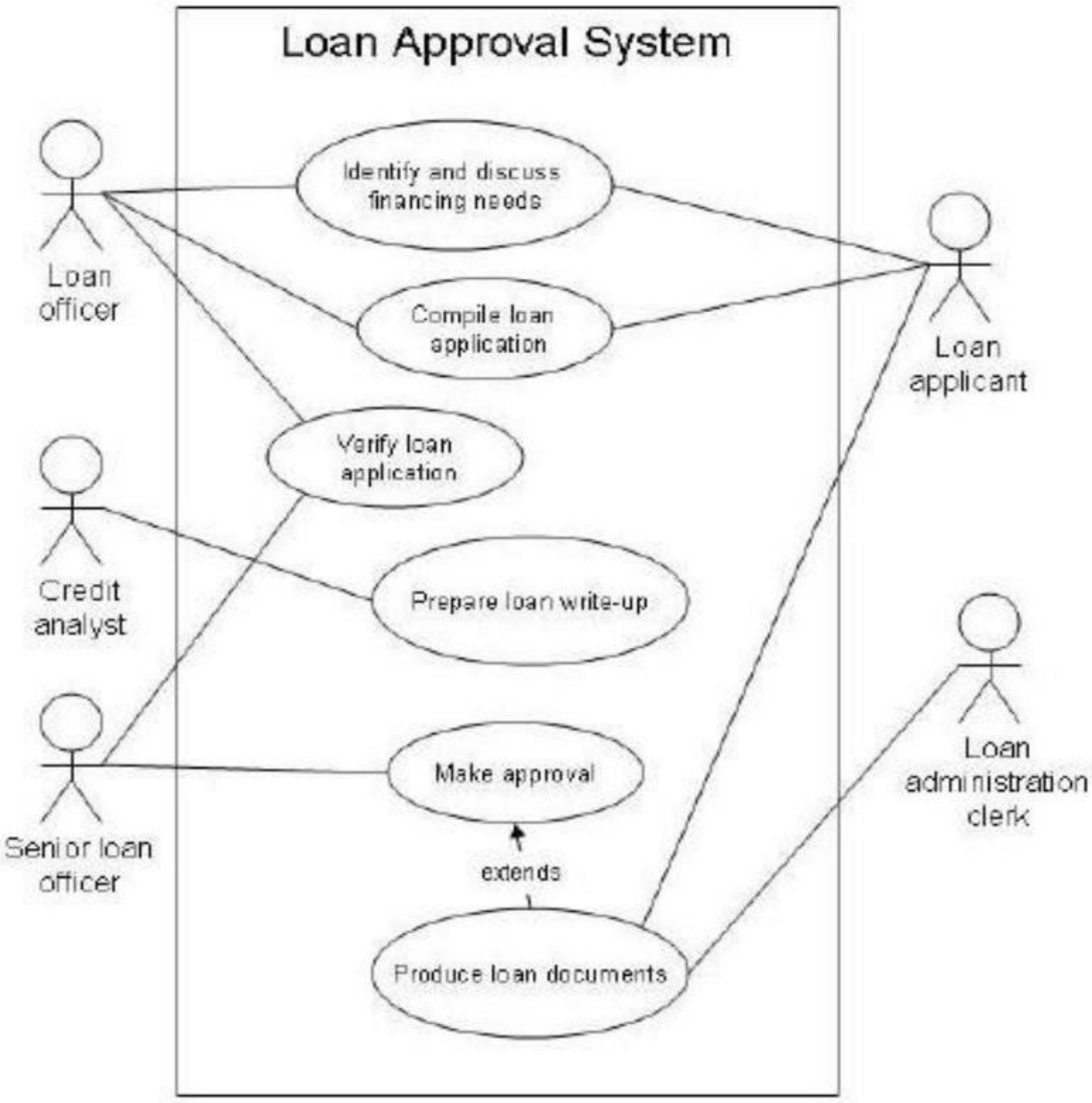
Logistic Regression:- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. Logistic Regression (LR) is a machine learning technique. The LR is very commonly used to solve binary classification problem. There are following basic postulation:

1. Binary logistic regression has binary dependent variables.
2. In binary regression dependent variables have level 1.
3. The included variables should have meaning. All included independent variables should be self-reliant.
4. The independent variables are related to the log odds linearly.
5. The sample size should be large for LR.

Python:- Python is an interpreter, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

1.3 System Design: -



.

CHAPTER-2
Literature Survey

The main objective of this paper is to predict whether the loan to particular person is safe or not. To predict loan safety, the SVM and Naïve bayes algorithm are used. First the data is cleaned so as to avoid the missing value in the data set. The main objective of this project is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four section (i) Data Collection (ii) Comparison of machine learning of Machine Learning models on collected data (iii) Training of system on most promising model (iv) Testing. The bank have to put in a lot of work to analyze or predict whether the customer can pay back the loan amount or not(defaulter or non -defaulter) in the given time .The aim of this paper is to find the nature or background or credibility of client that is applying for the loan .We use exploratory data analysis technique to deal with problem of approving or rejecting the loan request or in short loan prediction .The main focus of this paper is to determine whether the loan given to a particular person or an organization shall be approved or not.

A. Vaidya proposed a method for approval of loan prediction using logistic regression [1]. Logical regression is a machine learning technique which is very useful in prediction system. The approval of loan is a very important process in banking system.
A. Vaidya solves the problem by applying machine learning in a sample data set for loan approval applications. It also opens other areas on which machine learning is applicable.
A. Li and Q. Sun
[2] find a method to calculate risk involved in loan approvals for SMEs. A concept of loan consuming radius was introduced which was based upon supply chain in consumer market. F. M. Isik et al. develop a loan approval system using Business Process ExecutionLanguage BPEL [3]. The concept of BPEL is very useful in business firms. A reasoning engine was developed which removes some services from the BPEL process which are not necessary to complete a process. The system was applied on
machine learning approached in prediction systems. The machine learning approach was used for assessment of water quality. The paper concluded that machine learning is a very unimportant tool in prediction systems. C. Frank et al. [21] used machine learning in prediction of smoking status. Different machine learning approaches were applied and investigated for finding the smoking status. From the results its was ensured that logistic algorithm performs better. R. Lopes et al. applied machine learning approach for the prediction of credit recovery [22]. Credit recovery is very important issue for banking system. The prediction of credit recovery is a challenging task. Different machine learning approach was applied to predict the credit recovery and gradient expansion algorithms

(GBM) outperformed the other machine learning approaches.

After going through this literature it is found that loan approval prediction problem is very important for banking system. Machine learning algorithm are very useful in predicting outcomes even when data is very big in size. This paper investigated some machine learning algorithms and applied ML on test data set of loan approvals. Next section discussed the three machine learning approaches

CHAPTER-3
Proposed System

PROPOSED ALGORITHMS:

The following shows the pseudo code for the proposed loan prediction method.

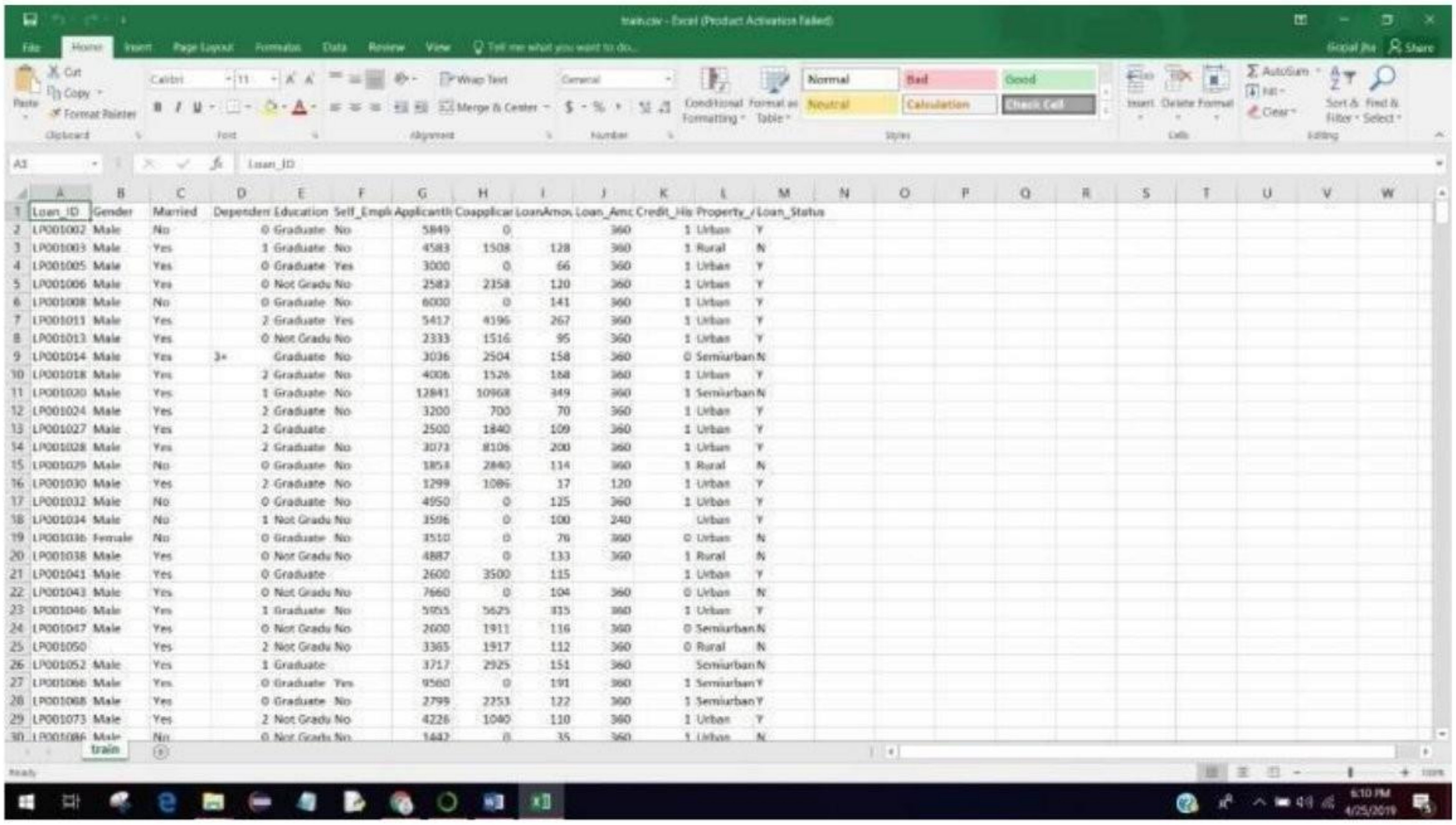
1. Loan the data
2. Determine the training and testing data
3. Data cleaning and pre-processing
   a) Fill the missing values with mean values regarding numerical values
   b) Fill the missing values with the model values regarding categorical variables.
   c) Outlier treatment
4. Apply the modeling for prediction
   a) Removing the load identifier
   b) Create the target variable (based on the requirement). In this approach, target variable is loan-status
   c) Create a dummy variable for categorical variable (if required) and split the training and testing data for validation.
   d) Apply the model-Decision Tree method, Logistic Regression method, Random Forest method.
5. Determine the accuracy followed by confusion matrix

CHAPTER-4

Result And Analysis

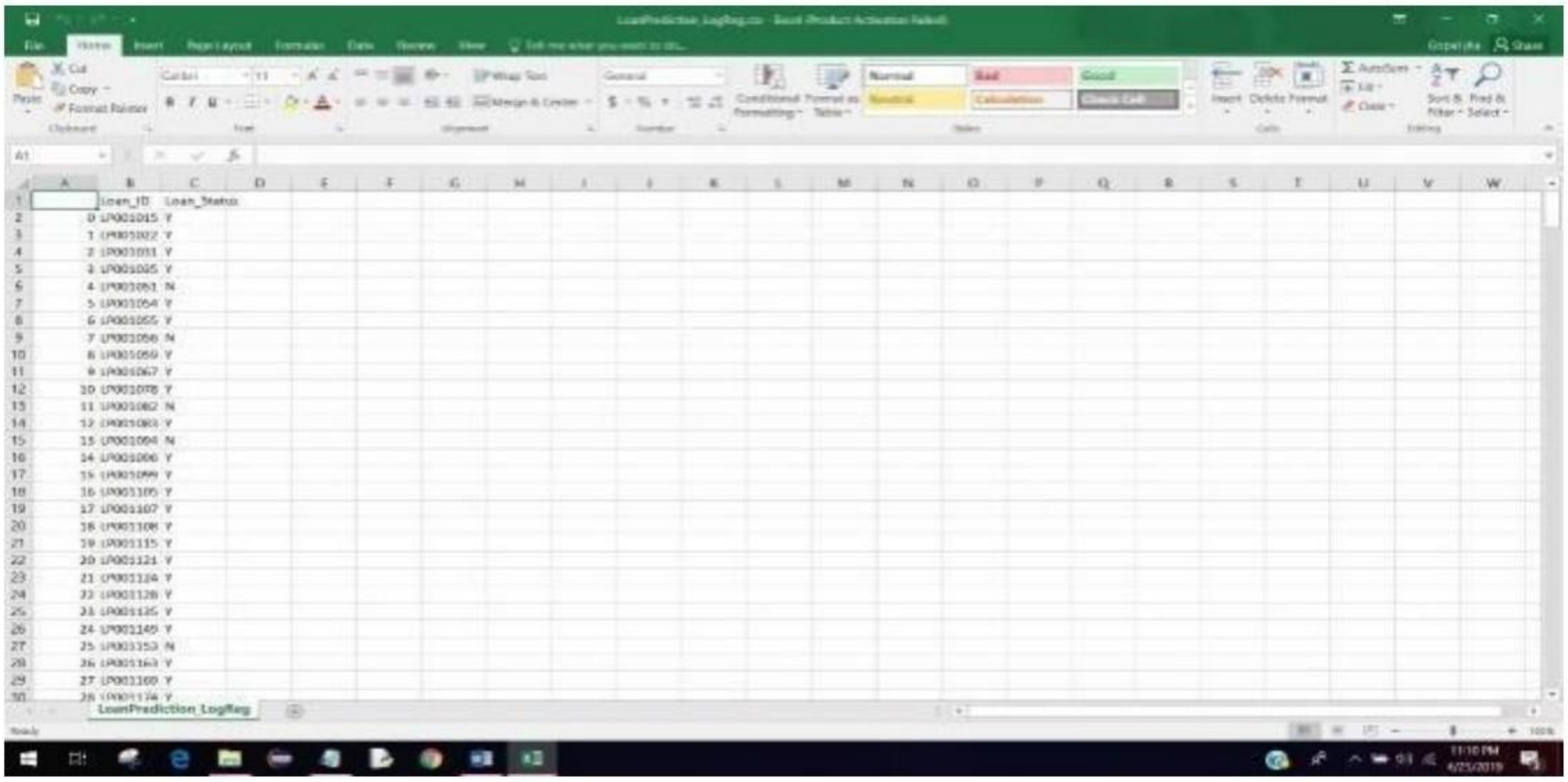
Three machine learning approaches are applied on the test data to predict the loan approvals of loan requests. Python programming language is used to implement machine learning algorithms. For training 70 percent data is used and 30 percent data is used for testing. The prediction accuracy of the different ML approaches is calculated and compared. The training data set is shown in figure 3.

On the basis of this train data set (shown in figure 3), system analyze rest of 30 percent data and predict the results in term of loan status either accepted or rejected. Results with loan status by applying the logistic regression (shown in figure-4(a)), decision tree (shown in figure-4(b)) and random forest (shown in figure-4(c)).Figure 5(a), 5(b) and 5(c) and 5(d) are demonstrating the histograms generated. Figure 5(a) is showing the histogram for applicant income. Figure 5(b) is showing histogram of co applicant income. Figure 5(c) is showing histogram of loan amount term. Figure 5(d) is showing histogram of loan amount.



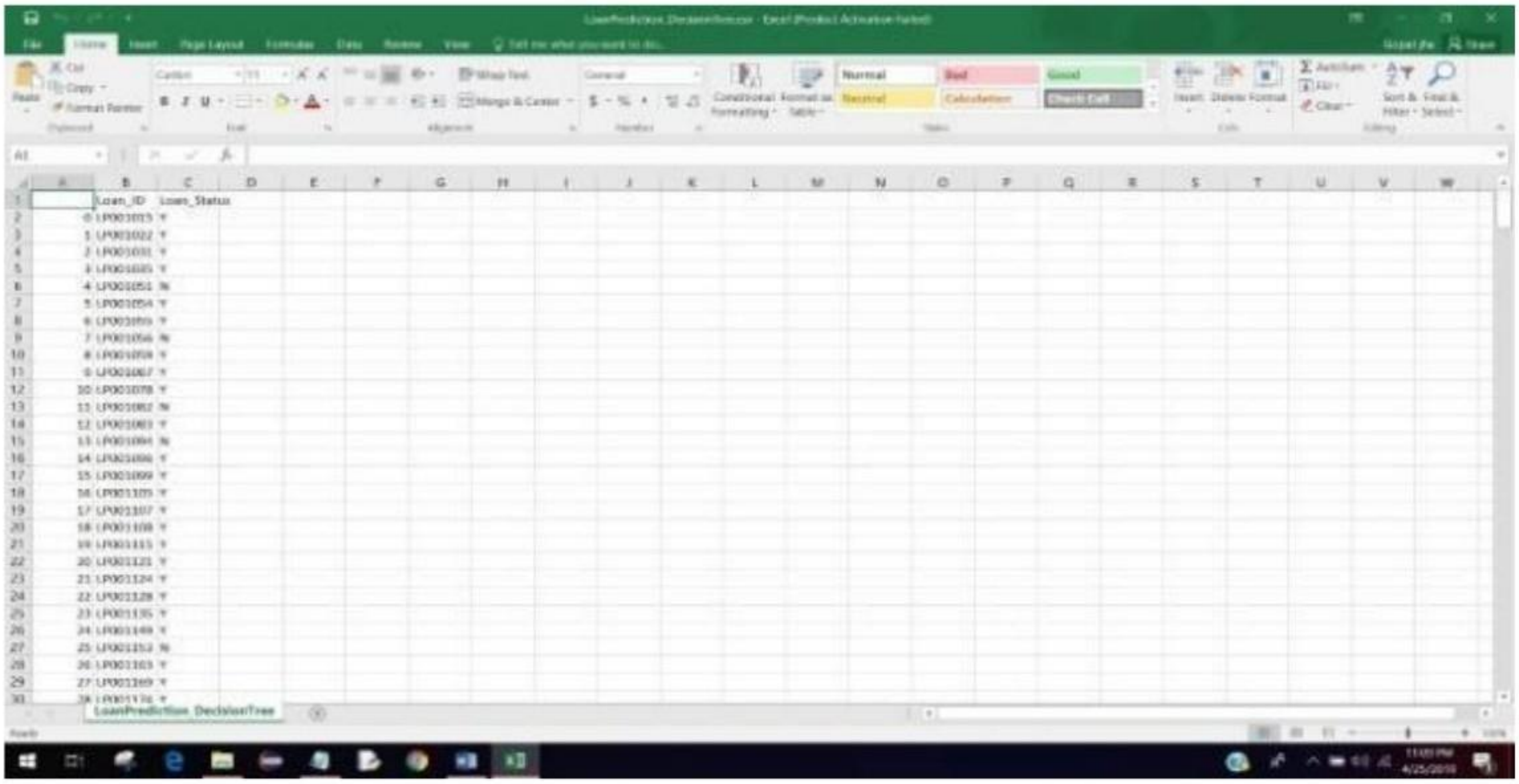
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Loan_ID	Gender	Married	Dependen	Education	Self_Empl	Applicantli	Coapplicar	LoanAmou	Loan_Amc	Credit_His	Property_/	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Gradu	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Gradu	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Gradu	No	3596	0	100	240		Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Gradu	No	4887	0	133	360	1	Rural	N
21	LP001041	Male	Yes	0	Graduate		2600	3500	115		1	Urban	Y
22	LP001043	Male	Yes	0	Not Gradu	No	7660	0	104	360	0	Urban	N
23	LP001046	Male	Yes	1	Graduate	No	5955	5625	315	360	1	Urban	Y
24	LP001047	Male	Yes	0	Not Gradu	No	2600	1911	116	360	0	Semiurban	N
25	LP001050		Yes	2	Not Gradu	No	3365	1917	112	360	0	Rural	N
26	LP001052	Male	Yes	1	Graduate		3717	2925	151	360		Semiurban	N
27	LP001066	Male	Yes	0	Graduate	Yes	9560	0	191	360	1	Semiurban	Y
28	LP001068	Male	Yes	0	Graduate	No	2799	2253	122	360	1	Semiurban	Y
29	LP001073	Male	Yes	2	Not Gradu	No	4226	1040	110	360	1	Urban	Y
30	LP001086	Male	No	0	Not Gradu	No	1442	0	35	360	1	Urban	N

Figure-3 Trained Data Set


LoanPrediction_LogReg.csv - Excel (Product Activation Failed)
LoanPrediction_LogReg

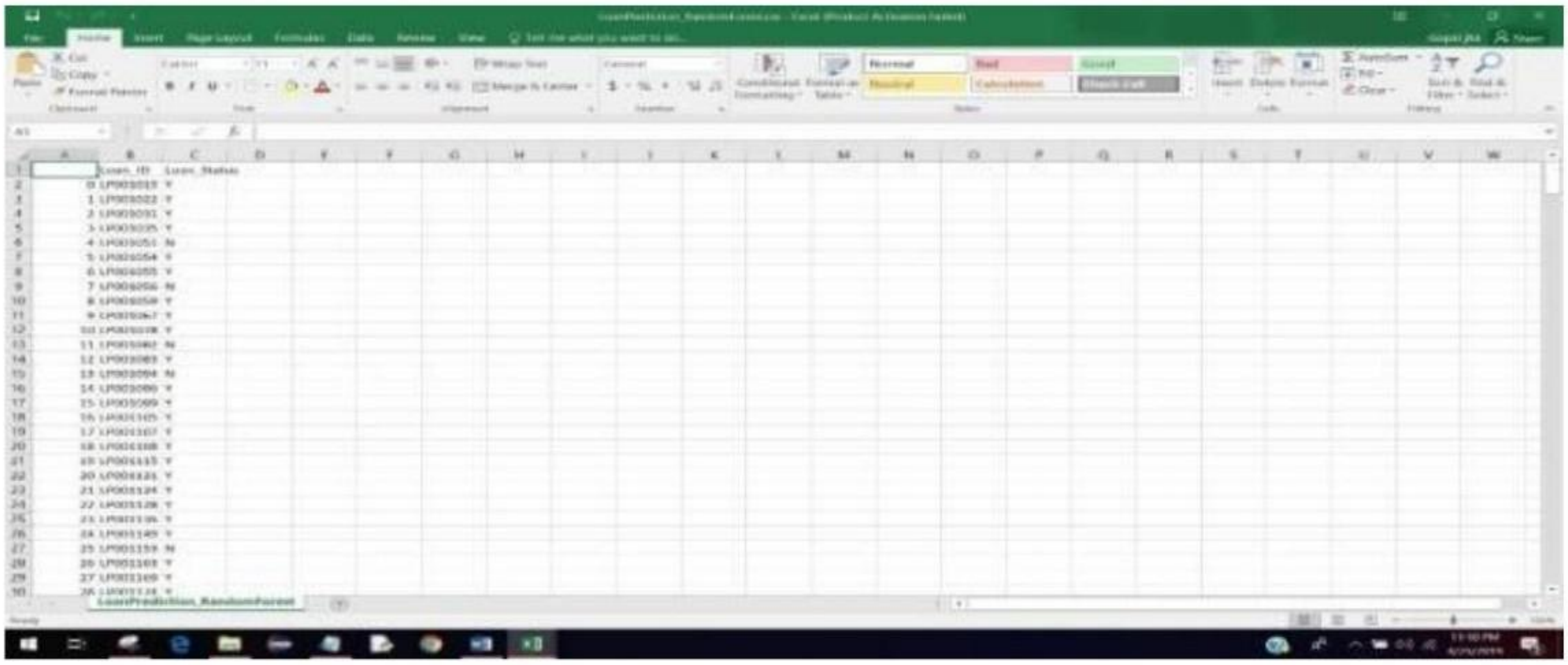
	Loan_ID	Loan_Status
0	LP001015	Y
1	LP001022	Y
2	LP001031	Y
3	LP001035	Y
4	LP001051	N
5	LP001054	Y
6	LP001055	Y
7	LP001056	N
8	LP001059	Y
9	LP001067	Y
10	LP001078	Y
11	LP001082	N
12	LP001083	Y
13	LP001094	N
14	LP001096	Y
15	LP001099	Y
16	LP001105	Y
17	LP001107	Y
18	LP001108	Y
19	LP001115	Y
20	LP001121	Y
21	LP001124	Y
22	LP001128	Y
23	LP001135	Y
24	LP001149	Y
25	LP001153	N
26	LP001163	Y
27	LP001169	Y
28	LP001174	Y

Figure-4(a): Logistics Regression Result with Loan Status


LoanPrediction_DecisionTree.csv - Excel (Product Activation Failed)
LoanPrediction_DecisionTree

	Loan_ID	Loan_Status
0	LP001015	Y
1	LP001022	Y
2	LP001031	Y
3	LP001035	Y
4	LP001051	N
5	LP001054	Y
6	LP001055	Y
7	LP001056	N
8	LP001059	Y
9	LP001067	Y
10	LP001078	Y
11	LP001082	N
12	LP001083	Y
13	LP001094	N
14	LP001096	Y
15	LP001099	Y
16	LP001105	Y
17	LP001107	Y
18	LP001108	Y
19	LP001115	Y
20	LP001121	Y
21	LP001124	Y
22	LP001128	Y
23	LP001135	Y
24	LP001149	Y
25	LP001153	N
26	LP001163	Y
27	LP001169	Y
28	LP001174	Y

Figure-4(b): Decision Tree Result with Loan Status.



	Loan_ID	Loan_Status
0	LP001015	Y
1	LP001022	Y
2	LP001031	Y
3	LP001035	Y
4	LP001051	N
5	LP001054	Y
6	LP001055	Y
7	LP001056	N
8	LP001059	Y
9	LP001067	Y
10	LP001078	Y
11	LP001082	N
12	LP001083	Y
13	LP001094	N
14	LP001096	Y
15	LP001099	Y
16	LP001105	Y
17	LP001107	Y
18	LP001108	Y
19	LP001115	Y
20	LP001121	Y
21	LP001124	Y
22	LP001128	Y
23	LP001135	Y
24	LP001149	Y
25	LP001153	N
26	LP001163	Y
27	LP001169	Y
28	LP001174	Y

Figure-4(c):Random forest Result with Loan Status.

Comparison analysis of prediction accuracy for three machine learning algorithms is shown in table -1.

Table-1: Comparison of prediction accuracy of machine learning algorithms

S.No	Machine Learning Algorithms	Prediction Accuracy Percentage
1	Logistic Regression	93.04
2	Decision Tree	95.0
3	Random Forest	92.53

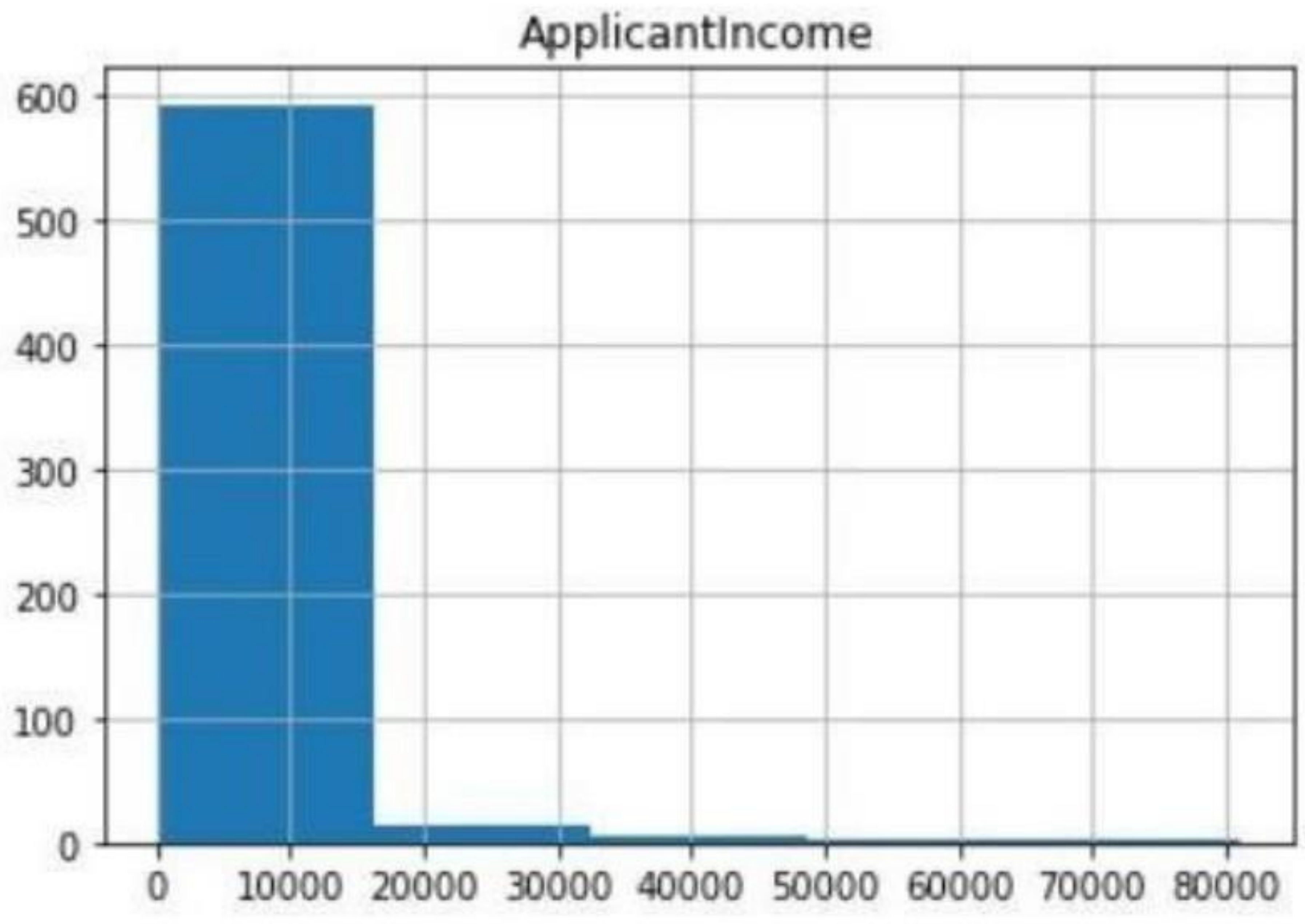


Figure-5(a): Histogram of Applicant income

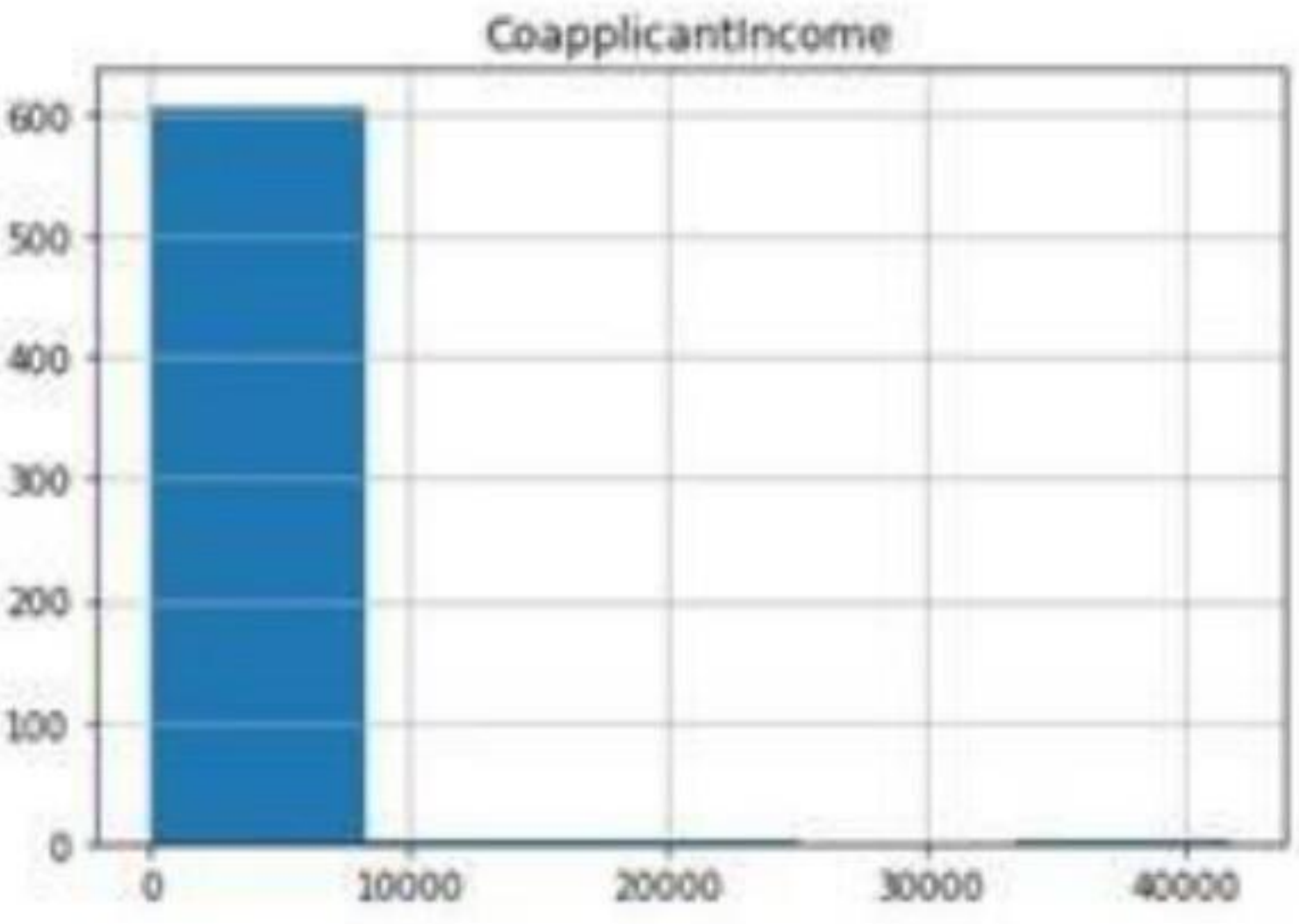


Figure-5(b): Histogram of Coapplicant income

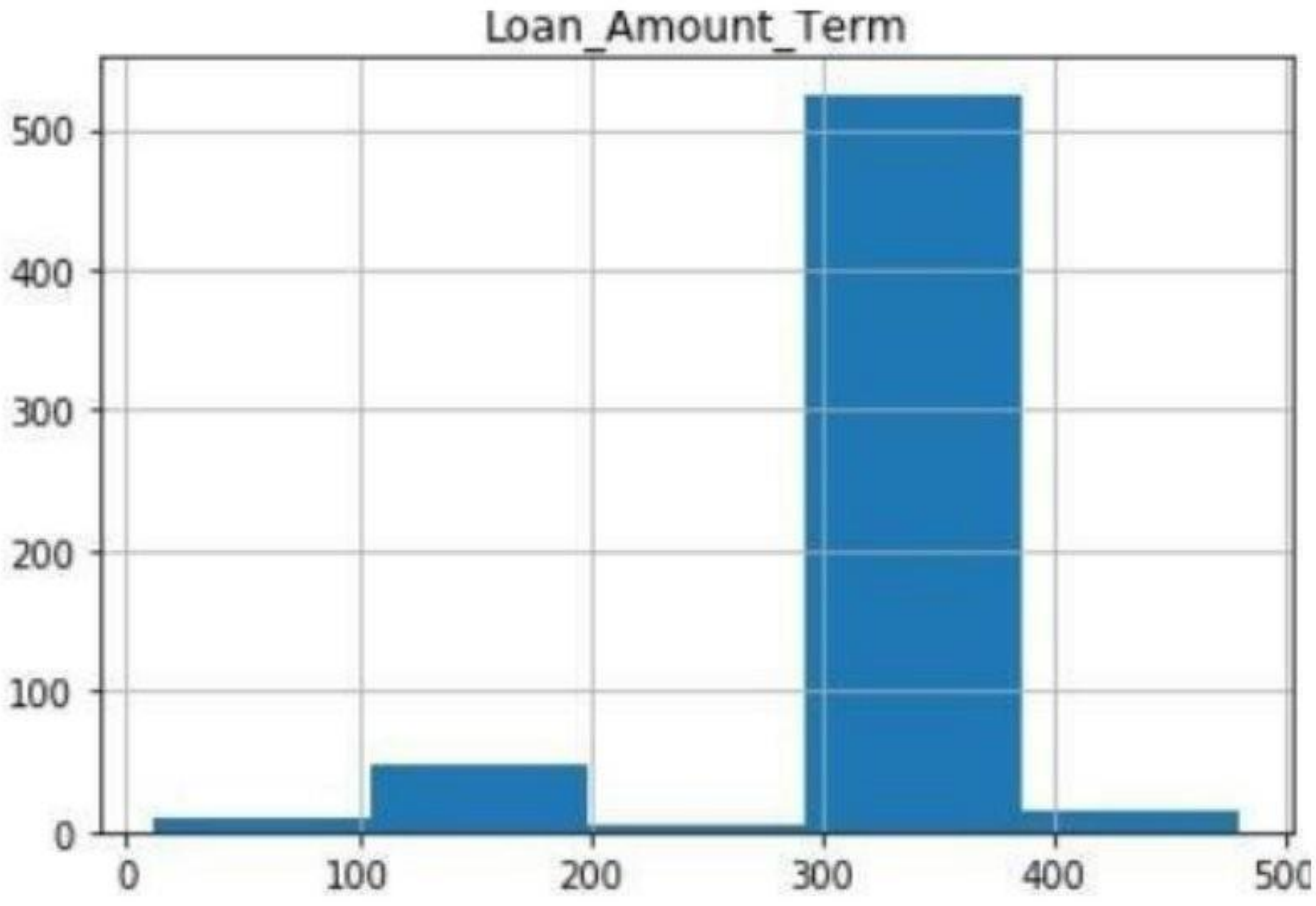


Figure-5(c): Histogram of Loan Amount Term

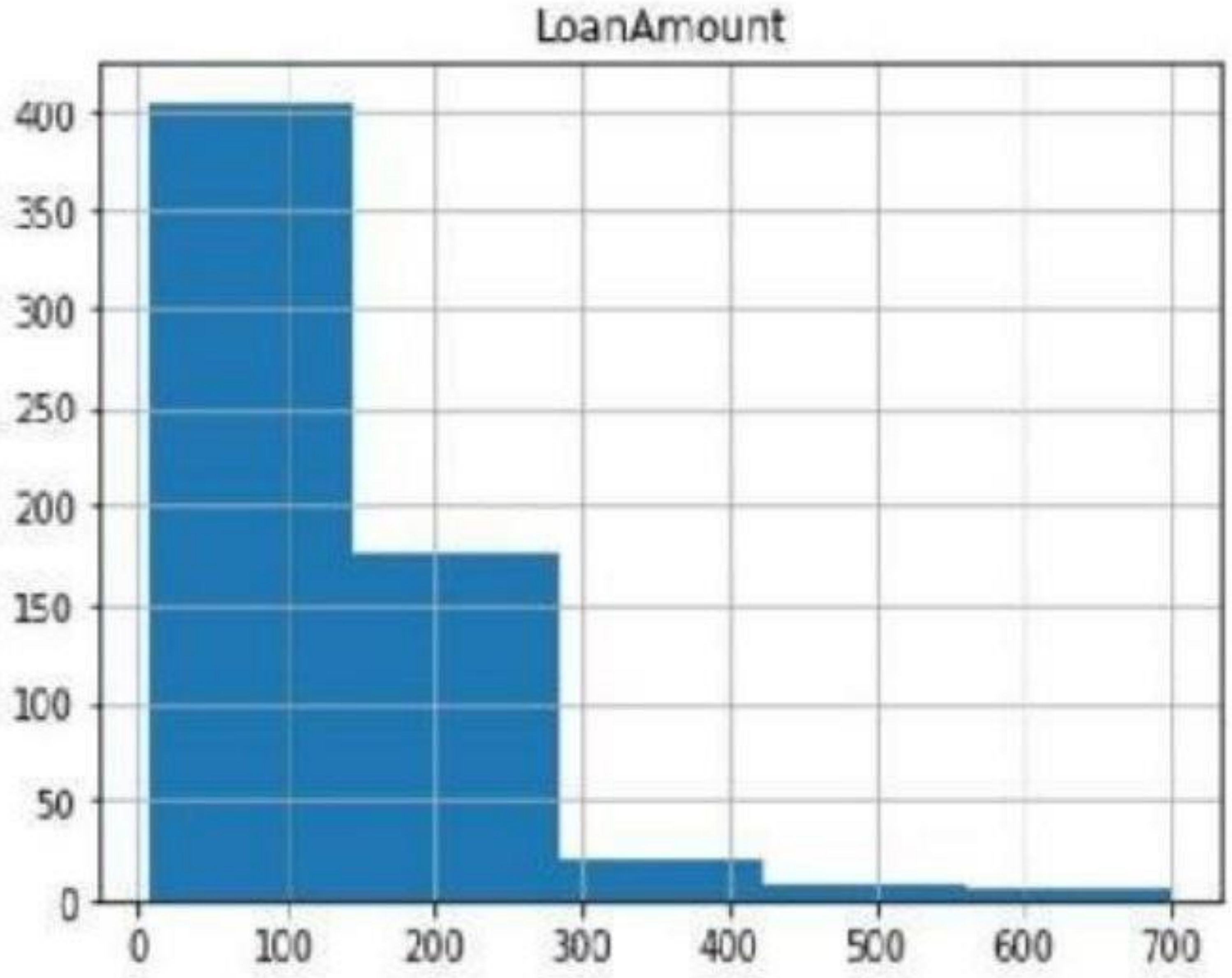


Figure-5(d): Histogram of Loan Amount

Chapter-5

5-Conclusion and Future Scope:-

This paper applied machine learning in prediction of loan approval. Three ML algorithms are used to predict the loan approval status of customers for bank loans. The results shown that the prediction accuracy is 93.04%, 95% and 92.53% for LR, DT algorithm algorithms respectively. Among three the accuracy of DT algorithm is best for prediction of loans. In future the Decision Tree algorithm can be applied on other data sets available for loan approvals to further investigate its accuracy. A rigorous analysis of other machine learning algorithms other than these three can also be done in future to investigate the power of machine learning algorithms for loan approval prediction.

References

[1] Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6.doi: 10.1109/ICCCNT.2017.8203946

[2] A.Li and Q. Sun, "The Risk of Loan Consuming by SMEs Based on the Supply Chain," 2012 International Conference on Management of e- Commerce and e-Government, Beijing, 2012, pp. 356-359.doi: 10.1109/ICMeCG.2012.24

[3] F. M. Isik, B. Tastan and P. Yolum, "Automatic Adaptation of BPEL Processes Using Semantic Rules: Design and Development of a Loan Approval System," 2007 IEEE 23rd International Conference on Data Engineering Workshop, Istanbul, 2007, pp. 944-951.doi: 10.1109/ICDEW.2007.4401089

[4] V. C. T. Chan et al., "Designing a Credit Approval System Using Web Services, BPEL, and AJAX," *2009 IEEE International Conference on e- Business Engineering*, Macau, 2009, pp. 287-294.doi: 10.1109/ICEBE.2009.46

[5] J. Lohokare, R. Dani and S. Sontakke, "Automated data collection for credit score calculation based on financial transactions and social media," *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*, Pune, 2017, pp. 134-138.doi: 10.1109/ETIICT.2017.7977024

[6] R. Yang, X. Zhou and W. Wang, "Is the Small and Medium-Sized Enterprises' Credit Default Behavior Affected by Their Owners' Credit Features?," *2011 International Conference on Management and Service Science*, Wuhan, 2011, pp. 1-4.doi: 10.1109/ICMSS.2011.5998460

[7] M. Bayraktar, M. S. Aktaş, O. Kalıpsız, O. Susuz and S. Bayracı, "Credit risk analysis with classification Restricted Boltzmann Machine," *2018 26th Signal Processing and Communications Applications Conference (SIU)*,
Izmir, 2018, pp. 1-4.doi: 10.1109/SIU.2018.8404397

[8] H. A. P. Pérez, J. A. P. Palacio and C. Lochmuller, "Fuzzy model Takagi Sugeno with structured evolution for determining consumer credit score," *2015 10th Iberian Conference on Information Systems and Technologies (CISTI)*, Aveiro, 2015, pp. 1-6.doi: 10.1109/CISTI.2015.7170485

[9] S. Yadav and S. Thakur, "Bank loan analysis using customer usage data: A big data approach using Hadoop," *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, Noida, 2017, pp. 1-8.doi: 10.1109/TEL-NET.2017.8343582