International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 05, Volume 8 (May 2021) https://www.ijirae.com/archives

PREDICTIVE ANALYSIS OF CLASSIFICATION ALGORITHMS ON BANKING DATA

Krishan Kumar Pandey
Computer Science and Engineering
Galgotias University, Uttar Pradesh, India
krishan.007305@gmail.com
Abhishek Giri
Computer Science and Engineering
Galgotias University, Uttar Pradesh, India
Abhishekgiri445@gmail.com
Saket Sharma
Computer Science and Engineering
Galgotias University, Uttar Pradesh, India
Saketsharma002@gmail.com
Anurag Singh
Computer Science and Engineering
Galgotias University, Uttar Pradesh, India
Anurag.singh485@gmail.com

Publication History
Research Article | Open Access
Peer-review: Double-blind Peer-reviewed
Article ID: IJIRAE/RS/Vol.08/Issue05/MYAE10084
Received: 07, May 2021
Accepted: 22, May 2021
Published Online: 08, June 2021
Volume 2021 | Article ID MYAE10084 | https://doi.org/10.26562/ijirae.2021.v0805.003
Krishnan,Abhishek,Saket,Anurag(2021).Predictive Analysis of Classification Algorithms on Banking Data. IJIRAE:: International Journal of Innovative Research in Advanced Engineering, Vol: VIII,99-106
doi: https://doi.org/10.26562/ijirae.2021.v0805.003
Editor-Chief: Dr.A.Arul Lawrence Selvakumar, Chief Editor, IJIRAE, AM Publications, India

Abstract: The entire human activity is generating about 2.5 quintillion bytes of data per day. It is growing exponentially. The tremendous volume of data is getting generated by different banking services, including real-estate loan, deposit accounts services, credit card issuing service, commercial loan lending and many more. The bank data is used for business analytic that keeps eyes on past growth also Predictive analysis opens doors for future growth. For this project, during the predictive analysis, we used some historical data generated by an organization during loan processing. We intended to predict whether a new applicant will have granted a loan or not. The analysis was done by building Machine Learning (ML) models using different classification algorithms like Random Forest(RF), Logistic Regression(LR), Support Vector Machine(SVM), Decision Tree(DT), Gradient Boosting, XG Boost. The classifier models were trained and tested on hyper parameters as the performance of each model was analyzed using Confusion Matrix. We compared efficiencies using Recall, Precision, Accuracy and F1 score in a tabular form among all models. LR yields an F1 score of 0.90 and 0.62 for class 1 (approved) and class 0 (not Approved), respectively. Also, XG Boost gave F1 score of 0.89 and 0.64 for the same class 1 and 0.

Keywords: Loan Data, Machine Learning, Classification Algorithms, F1 Score.

I. INTRODUCTION

Industry 4.0 is automating the entire human activity with the help of Artificial Intelligence (AI), Machine Learning (ML), Cloud Computing and IoT devices [1]. Automation is pushing the industries to a new level includes banking. The services provided by banks are producing a variety of data in enormous volume with high velocity [2]. Lending a loan is one of the services which increase more liquidity in the market. Excessive cash flow means more selling and buying of goods which is the building block of the economy. In recent times, we have seen a lack of cash flow due to the covid-19 pandemic and contract the Indian economy by 23.9% in the April-June Quarter of 2020. Although approving a loan is not as smooth as we think.

During the process, we witness issues on either end either it is the lender or borrower. The borrower must have a good credit score and repayment of the previous loan on time. If you are a fresh applicant, your profile should be strong enough. From the lender's perspective, the default risk should be low, lending money makes a profit or not? are the final questions for lending organizations [3]. During the entire process, the bank examines the applicant's profile manually on the various factors and most of them are repeatedly used for the upcoming borrowers. Therefore, loan processing consumes time. We can prudently use historical data to avoid the rigorous process for upcoming applications with the help of machine learning (ML). ML is widely used in various industries and making future predictions which are business advantages for the industry. ML can play a vital role to predict genuine and low-risk customer for a loan with the help of previous bank data. In this project, we used python as a programming language and open-source software, Jupyter Notebook platform. Python has inbuilt libraries to perform ML pipelines on the dataset. Some of the libraries are pandas, NumPy, Sk-learn, Matplotlib and seaborn.

II. RELATED WORK

We considered various research papers for the predictive analysis and summarized them based on their predictive analogy. They proposed models built on several classification algorithms. The threshold of acceptance and rejection of applications was different in each model. Ashlesha Vaidya [1] discussed a probabilistic model to categorize the loan seekers whether their loan application is approved or not. The research paper drew a mathematical outline using logistic regression that is good in binary classification problems. In the findings, if the probability was above 0.5, approve the loan else reject the application. K. Gana Sai Prasad, S., Chidvilas, P.V., and Kumar, V.V.[3] used two different supervised learning classification algorithms, Random Forest and SVM as the loan default prediction model. The model's predictive analysis on the data set was using a confusion matrix. The accuracy result concluded as if it is greater than 75 %, the borrower is more likely to pay the loan back. Since the default risk is low, therefore grant loan otherwise rejects the application. Chaudhary, S.Baliyan, and B.Katheria, Y.[4] proposed a predictive system built on classification models. Two Tree-based classifiers are Decision Tree and Random Forest. They were trained and tested on the historical bank data and later predicted. The efficiency attained after hyperparameter tuning was 85% and 85.3 % in Decision Tree and Random Forest, respectively.

III. PREDICTIVE ANALYSIS METHODOLOGY

A. Data Wrangling

The data preparation technique is involved in this segment of predictive analysis. Data wrangling transform data to do Exploratory Data Analysis (EDA) and the processed data is transferred to machine learning algorithms for the prediction.

1) Data Gathering: The raw data is gathered from an open source repository known as Kaggle. In the gathered data there were various features that consumer had to fill while applying for a personal loan and based on that either they were given a loan or rejected. The data had 19 such independent features that were filled by consumers and one dependent feature mentioned as CPL_Status (Yes/No), telling whether the consumer was given a loan or not. The other important attributes are Credit History (score of the previous loan), Income of applicant (App_Income_1) and co-applicant (App_Income_2), for how long the consumer seeks a loan (CPL_Term) and how much amount one needs (CPL_amount), self-employed or not (SE), etc [1]. The other attributes can be found in the Table 1 given below.

Table 1. Columns in the dataset

Column Name	
Loanapp_ID	App_Income_1
Sex	App_Income_2
Marital_Status	CPL_Amount
first_name	CPL_Term
last_name	Credit_His
email	Prop_Area
address	INT_ID
Dependents	Prev_ID
Qual_var	AGT_ID
SE	CPL_Status

2) Data Assessing: The bank data assessment is to get answers to questions like How many columns and rows are there? What different data types are present? Is there any missing value? Are there any issues with data like tidiness and quality issues? The assessment will perform in two ways: visual and programmatic. In visual evaluation, we use python code to get the first five rows of the data set. Similarly, the last five rows can be another way of accessing data. But, this method is not feasible for finding the total number of missing values in each column.

If data have more columns, it doesn't show all columns at a time. So, in this scenario, we use a programmatic assessment of data to fetch all the information. Afterwards, We get a crucial description of the dataset having the number of rows and columns, missing values in each column, the data type of each attribute.

3) Data Cleaning: We have seen issues in the raw data having missing values in various columns, merged two columns app_income_1 and app_income_2 as Total_income. We removed few attributes not required in further analysis. We make a copy of the original data named 'df_copy before dealing with the missing values because it doesn't affect the content in the original dataset. After filling in the missing values, we removed few attributes as well. Now we can observe the changes made before and after the data wrangling process. Clearly, in Fig. 2, we can visualize that no more missing values are present. It is ready for further analysis where Fig. 1 shows the frequency of missing values present in the dataset.

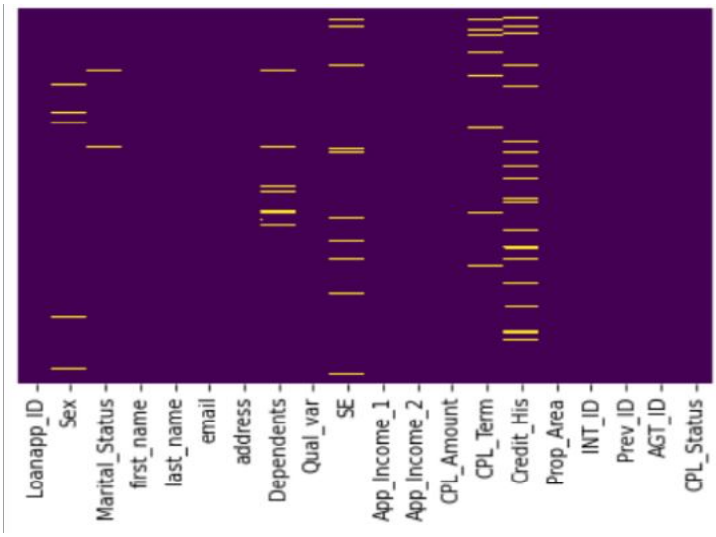

Fig. 1. Dataset before Data Wrangling

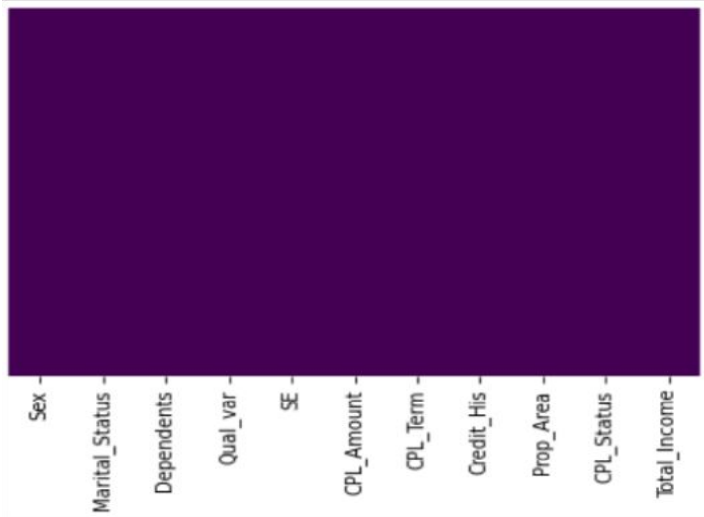

Fig. 2. Dataset after Data Wrangling

B. Exploratory Data Analysis (EDA)

EDA is the process of getting explanatory answers to various presumed questions such as the relationship between attributes to find out the anomalies, the statistical summary of correlated features etc. All these questions can answer using graphs with a proper correlation among them. The visualization will use Univariate analysis, Bivariate analysis and Multivariate analysis. EDA is performed on the cleaned dataset in Jupyter Notebook using two visualization libraries in python, matplotlib and seaborn. Univariate analysis can perform on a single attribute where visualization depends upon the data type of features. Examples concerning bank data are present in Fig. 3 and Fig. 4.

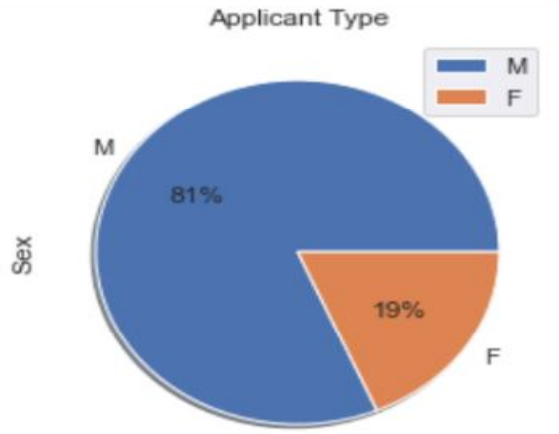

Fig. 3. Pie chart of applicant type

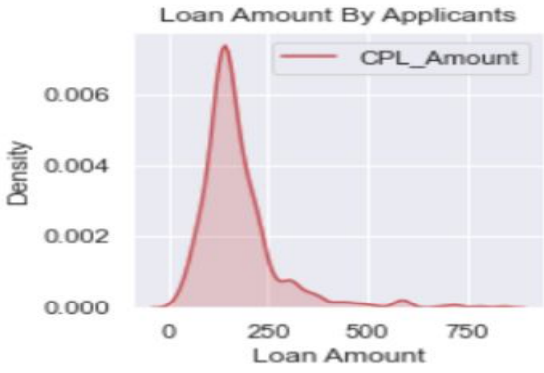

Fig. 4. Density distribution of loan amount

The bivariate analysis uses two different attributes; either it is categorical or continuous data. The bivariate visualization is in Fig. 5 and Fig. 6.

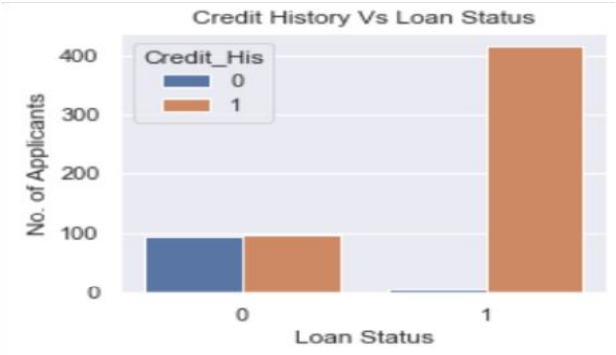

Fig. 5. Credit history vs loan status

Fig. 6. Loan Status vs Total Income

Similarly, for multivariate analysis, we use more than two attributes to visualize together depending on the data type. One can smoothly draw the relationship between credit history, number of 'dependents' and 'loan status' in Fig.7.

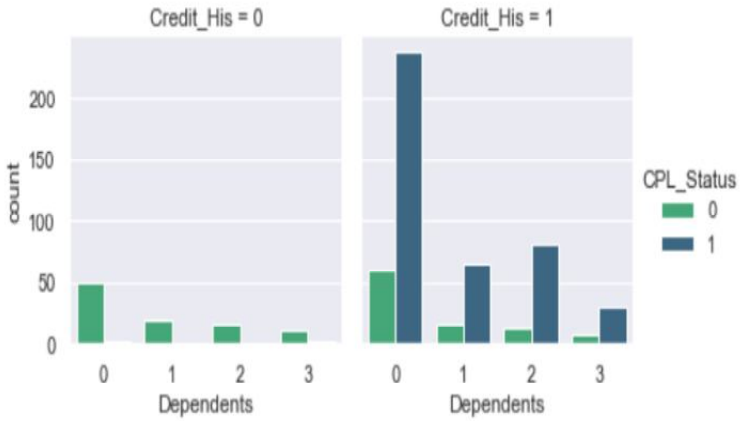

Fig. 7. Credit History Vs Dependents Vs Loan Status

C. Model Creation

The dataset gets cleaned and analyzed in the previous sections. Here, we will be splitting the dataset in the 7:3 ratio as training and testing data, respectively and parsing it into different classification algorithms. The model creation architecture has drawn in Fig 8.

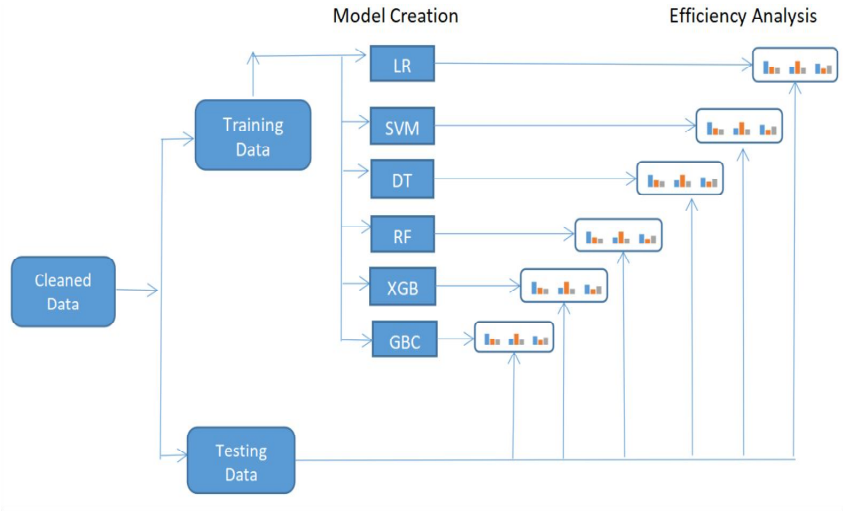

Fig 8: Model building architecture

Few categorical attributes to be converted into dummy variables such as 'Sex', 'Marital_Status', 'Qual_var', 'SE', 'Prop_Area' and 'Dependents'. After converting into dummy variables, models get trained and noted down the accuracy of each model on default parameters for the training dataset shown in Fig. 9.

	model	best_score	best_params
0	SVM	0.662025	{}
1	Random_Forest	0.790315	{}
2	Logistic_Regression	0.775978	{}
3	Decision_Tree	0.699261	{}
4	Gradient_Boosting	0.797237	{}
5	XGBoost	0.769193	{}

Fig. 9. Best score on default parameters

The tuning of hyperparameters on training data for each model to attain the best performance up to a certain level so that models don't overfit [4]. Accuracy score after hyperparameter tuning along with the recommended hyperparameter values. A resampling method k-fold cross-validation was used with the k value as 5 in each algorithm shown in Fig. 10.

	model	best_score	best_params
0	SVM	0.797182	{'kernel': 'linear'}
1	Random_Forest	0.827551	{'max_depth': 3, 'n_estimators': 5}
2	Logistic_Regression	0.822873	{'C': 0.1, 'penalty': 'none', 'solver': 'newto...
3	Decision_Tree	0.825198	{'criterion': 'gini', 'max_depth': 5, 'max_fea...
4	Gradient_Boosting	0.813516	{'learning_rate': 0.05, 'n_estimators': 10}
5	XGBoost	0.829850	{'learning_rate': 0.15, 'max_depth': 2, 'n_est...

Fig. 10. Best score on hyperparameters

IV. EXPERIMENTAL RESULTS

Since the problem statement "to predict whether a borrower granted a loan or not" is a binary classification problem and fully business-oriented in the banking organizations. Therefore, our aim should be to develop such a model that doesn't miss out on the genuine customer for the loan. So, in such type of business scenario, we can't rely on the accuracy score of any model alone. Use of other units of confusion matrix shown in Fig. 11, such as recall, precision, f1 score, etc become crucial to analyze the result concerning the problem statement [5] [6]. Performance metrics derived from confusion matrix using Fig. 11 :

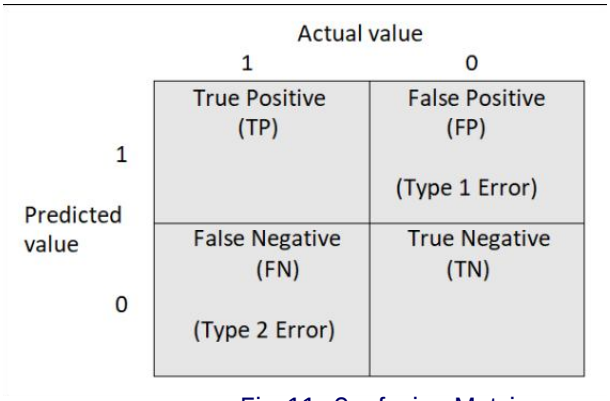

Fig. 11. Confusion Matrix

a) Accuracy: The fraction of true values predicted by model is called accuracy. Its range lies between 0 to 1.
Accuracy = (TP + TN) / (TP + FP + FN +TN)

b) Recall (True Positive Rate): Basically it says, out of actual positive values how many fraction we predicted positive. Recall = TP / (TP + FN)

c) Precision (Positive Prediction Value): By definition, out of total predicted positive values how many fractions were actually positive. Precision= TP / (TP + FP)

d) F1 Score: It is the harmonic mean of Recall and Precision. Using F1 score we measure Recall and Precision all together. F1 Score = (2* Precision * Recall) / (Precision + Recall)

When banks approve the loan, they would never like to lose the person who deserves the loan. If such clients missed out on the ML models, there would be a business loss to the organizations. Similarly, if they grant the loan to an undeserving candidate, there is also a possibility of losing money. So, In this scenario, both class 1 (Approved) and class 0 (Not Approved) become very useful. Therefore, predictive analysis using the F1 score becomes more relevant for this particular problem. Although, we also calculate other performance metrics for each model to have better intuition.

The measurement of performance metrics of each classification algorithm on test data is in the tabular format from Fig. 12 to Fig. 17. Logistic regression and XG Boost has shown in Fig. 12 and Fig. 17, respectively. The plot seems to have a better F1 Score for each class (1 or 0) among all the classifiers.

Classification report of Logistic Regression :

	precision	recall	f1-score
1	0.83	0.99	0.90
0	0.92	0.47	0.62
accuracy			0.84
macro avg	0.88	0.73	0.76
weighted avg	0.86	0.84	0.82

Fig. 12. Logistic Regression

Classification report of Decision Tree :

	precision	recall	f1-score
1	0.84	0.96	0.89
0	0.81	0.51	0.63
accuracy			0.83
macro avg	0.82	0.73	0.76
weighted avg	0.83	0.83	0.82

Fig. 13. Decision Tree

Classification report of Gradient Boosting :

	precision	recall	f1-score
1	0.82	0.96	0.89
0	0.82	0.45	0.58
accuracy			0.82
macro avg	0.82	0.71	0.73
weighted avg	0.82	0.82	0.80

Fig. 14. Gradient Boosting

Classification report of Random Forest :

	precision	recall	f1-score
1	0.83	0.97	0.89
0	0.86	0.47	0.61
accuracy			0.83
macro avg	0.84	0.72	0.75
weighted avg	0.84	0.83	0.81

Fig. 15. Random Forest

Classification report of SVM :

	precision	recall	f1-score
1	0.82	0.99	0.90
0	0.96	0.43	0.59
accuracy			0.84
macro avg	0.89	0.71	0.75
weighted avg	0.86	0.84	0.81

Fig. 16. SVM

Classification report of XGBoost :

	precision	recall	f1-score
1	0.84	0.93	0.89
0	0.76	0.55	0.64
accuracy			0.83
macro avg	0.80	0.74	0.76
weighted avg	0.82	0.83	0.82

Fig. 17. XGBoost

V. CONCLUSION AND FUTURE SCOPE

In this research paper, we perform predictive analysis of various classification algorithms based on the F1 score and other performance metrics as well. Logistic Regression classifier and XGBoost classifier stand out to be better ML models based on F1 score. The efficiency of LR to be an F1 score of 0.90 and 0.62 for class 1 and class 0, respectively. Also, XGBoost has an F1 score of 0.89 and 0.64 for the same class 1 and 0. Since the banking firms produce dissimilar data during the process of different type of loan services, accept only a few repeated features. So, we can't generalize one method for all kind of data produced. Therefore, a broad scope of enhancement is still needed while enforcing machine learning in the banking sector. Eventually, the default risk of customers would be a big challenge for the lenders.

ACKNOWLEDGMENT

We would like to express our deep gratitude to guide Mr. Anurag Singh, our supervisors, for their patient guidance, enthusiastic encouragement and useful critics of this work. We would also like to thank them for their suggestions and support in our progress.

REFERENCES

1. A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval", 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 2017, pp. 1-6, https://doi:10.1109/ICCCNT.2017.8203946.
2. P. S. Patil and N. V. Dharwadkar, "Analysis of banking data using machine learning," International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2017, pp.876-881,doi:10.1109/I-SMAC.2017.8058305.
3. Gana, K., Prasad, S., Chidvilas, P.V., & Kumar, V.V. "Customer Loan Approval Classification by Supervised Learning Model", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.
4. Chaudhary,S., Baliyan,B., Katheria, Y., "Loan Prediction System Using Decision Tree and Random Forest Algorithms", International Journal of Emerging Technology and Innovative Engineering Volume 6, Issue 06, June 2020 (ISSN: 2394 – 6598).
5. Amornsamankul, S., Pimpunchat, B.,Triampo,W.,Charoenpong, J., & Nuttavut, N., "A Comparison of Machine Learning Algorithms and Their Applications", International journal of simulation: systems, science and technology. https://DOI:/10.5013/IJSSST.a.20.04.08.
6. R.Ramesh, "Predictive analytics for banking userdata using AWS Machine Learning cloud service", 2nd International Conference on Computing and Communications Technologies (ICCCT), Chennai, India, 2017, pp. 210-215, https://doi:10.1109/ICCCT2.2017.7972282
7. Gupta, A., Pant, V., Kumar, S., & Bansal, P.K., "Bank Loan Prediction System using Machine Learning". 9th International Conference System Modeling and Advancement in Research Trends (SMART), 2020,423-426.
8. Quinlan, J., "Induction of Decision Trees ", Machine Learning 1: 81-106, 1986. Chotwani, P., Tiwari, A., & Hooda, M.
9. "Fraudulent Loan Prediction using Machine Learning Algorithms", Indian Journal of Public Health Research and Development, 2020, 10, 845-850.
10. Supriya, P., Pavani, M., Saisushma, N., Kumari, N.V., & Vikas, K. "Loan Prediction by using Machine Learning Models", 2019
11. M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 490-494, https://doi:10.1109/ICESC48915.2020.9155614
12. B., Reddy, C.K., Srinivas, C.K., & Reddy, K.L. "Loan Delinquency Prediction using Machine Learning Technique", 2020.