ISSN No:-2456-2165

An Exhaustive Investigation on Loan Prediction in Banks using LRD

Himanshi Sharma[1],
Ishika Tyagi[2],
Gauri Agarwal[3],
Deeksha Gupta[4]

$Student^{1-4}$
Department of Computer Science and Engineering, IMS Engineering College$^{1-4}$,
AKTU, Ghaziabad, Uttar Pradesh – 201015

Abstract:- Although there are various items that banking systems can sell to make money, their primary source of income is their credit card system. Now that the banking industry is doing better, but because banks only have so many assets to lend to, choosing who will be a safer option for the bank and to whom the loan can be provided is usually a procedure. The banking sector still desires a more rigorous predictive modelling framework for a number of issues. Predicting loan defaulters is a difficult task for the banking sector. The loan status, which is the first stage of the loan lending procedure, is one of the quality metrics of the loan. Using machine learning, it is possible to automate the process of determining whether a loan should be authorized or not to the loan asker. This is done in more efficient way by searching through available data for prior loan recipients, after which machine learning methods are used to train the system based on the histories and experiences on available data. There are several ways to analyze the former mentioned issues on loan prediction in accordance with the research conducted by many researchers in this era. In this research paper we basically conducted the exhaustive investigation on DGHI dataset for analyzing the customer eligibility whether he is eligible for loan or not using LRD machine learning algorithms (i.e. Logistic Regression, Random Forest and Decision Trees). The experimental study conducted has been divided into two phases: Training and Testing of the available data. On the basis of investigation conducted we decided to choose Logistic Regression as the best technique for probability of loan prediction for the customer. The results obtained and selection of Logistic Regression as the suitable technique for the given approach has been done on the basis of parameters such as: Loan_id, Gender, Married, Education, Self-employed and so on. For future work it has been decided to improve the accuracy and precision of Logistic Regression.

Keywords: *Decision Tree, Logistic Regression, Machine Learning algorithms, Medical Insurance, Prediction, Random Forest.*

I. INTRODUCTION

Around the world, banks, housing finance firms, and some NBFC companies facing difficulties in providing loans on several services such as home loans, personal loans, business loans and so on [1]. According to the research, former mentioned loan services can be automated by using existing machine learning techniques in predicting the probability of a specific person whether he is eligible for loan or not on the basis of the histories available in the dataset of bank [2]. Apart of artificial intelligence [3], machine learning enables a computer program to automatically learn from a previous history on the basis of several recommendation algorithms [4] that helps in analyzing data, spotting trends, and incorporating the least amount of human interaction possible. Machine learning tools can perform almost any task that can be finished with a data-defined pattern or set of rules. This enables businesses to alter procedures that were previously only adaptable by humans responding to inquiries from customers, bookkeeping, and evaluations. The distribution of loans is generally always a bank's main business. The majority of the bank's assets are directly linked to the income the bank receives from the loans it makes. The primary goal of the banking industry is to ensure that customers' money is secure. The acceptance of loans is a crucial stage for banking firms. Despite the thorough screening and validation process, there is no guarantee that loans provided by various banks and financial institutions today will be returned. The chosen individual deserves the job the most out of all the applicants. The decision to approve a loan for an application used to be made manually and took a long time. However, it is now possible to use machine learning (ML) techniques [5] to predict whether or not a loan application will be granted. Machine learning algorithms are very useful for making predictions when evaluating massive amounts of data. Machine learning specifically supports data prediction, data decision-making, and data learning from the available and on the basis of customer history/activities. The behavioral scoring method is used to determine new clients based on their payment history and personal information [6]. According to previous research numerous machine learning algorithms for

predicting the loan by the customer has been implemented such as Decision Trees, Logistic Regression, Random Forest Classifiers, SVM [7]. Different algorithms yielded differing degrees of accuracy depending on the dataset provided. In the mentioned research work, we primarily conducted the exhaustive investigation on the available dataset "DGHI" [8] using the LRD techniques namely Logistic Regression, Random Forest and Decision Tress respectively. The prediction analysis has been conducted on the basis of mentioned parameters such as Loan_id, Gender, Married, Education, Self-employed and so on.

The paper has been organized as follows: Section 2, basically describes the literature survey conducted by the numerous researchers on the former mentioned objective. Section 3, explains the implementation work in the terms of related work presented of the proposed model used and conducted on the basis of numerous parameters (Loan_id, Gender, Married, Dependents, Education, Self_employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area and Loan_Status) on the available dataset "DGHI". Section 4, explains the results and observations obtained on the basis of the investigation conducted on the Logistic Regression, Random Forest and Decision Tress respectively presented in the terms of Heat map, comparative analysis conducted on the parameters principle, accuracy, inference parameters, advantages and disadvantages respectively. Section 5, explains the future work for further improvement in the proposed model.

II. LITERATURE SURVEY

In [9] Madaan et.al. a thorough and comparative analysis of two algorithms—Random Forest and Decision Trees. The identical dataset was utilized for both methods, and analysis of the findings revealed that the Random Forest algorithm performed better than the Decision Tree technique and did so with a far higher degree of accuracy. We received an accuracy of 80% from the Random Forest Classifier and an accuracy of 73% from the Decision Tree approach. As a result, it appears that the Random Forest model is a better choice for this type of data. This study explored, analyzed, and developed a machine learning algorithm to accurately determine whether a person, given a set of characteristics, is likely to default on a loan. Lending Club may utilize this kind of model to detect certain financial characteristics of prospective borrowers who may be at risk of defaulting and failing to repay their loans by the deadline.

In [10] Tejaswini et.al. presents a thorough study in the forecast on an approval for loan request by customer on the basis of three machine learning algorithms: Logistic Regression, Decision Tree (DT), and Random Forest (RF). According to the experimental findings, the Decision Tree machine learning algorithm is more accurate than Logistic Regression and Random Forest machine learning methods. This component is simple to plug into numerous other systems. Numerous instances of computer faults, content mistakes, and most significant weight of features are rectified in automated prediction system. This application is operationally sound and complies with all Banker specifications. This component is simple to plug into numerous other systems.

In [11] Dhanashri et.al. presents a study on the basis of performance metrics like sensitivity and specificity. The final outcomes have demonstrated that the model produces varied outcomes. The model is slightly superior because it takes into account variables (consumer characteristics such as age, purpose, credit history, credit amount, credit duration, income, etc.) in addition to checking amount information (which indicates a customer's wealth) that must be taken into account to accurately calculate the probability of default loan. Therefore, by using a suitable approach, it is simple to identify the right customers to be targeted for loan grants by evaluating their chance of loan neglect. The financial sector can benefit from data mining strategies for better focusing on and attracting new customers, most important client maintenance, programmed credit endorsement which is used to prevent extortion, misrepresentation identification over time, providing section-based products, client investigation, exchange designs over time for better maintenance and relationships, risk the executives, and advertising.

In [12] Ugochukwu et.al. presents a research that shows good performance accuracy, with the Random forest algorithm scoring the best (95.55%) and the Logistic regression approach scoring the lowest (80%). In terms of precision-recall and accuracy, our models fared better than two of the three loan prediction models that were discovered in the literature. The techniques and algorithms used in this study could, in general, help financial authorities, businesses, and individual borrowers succeed in their efforts to streamline the loan approval process.

In [13] Bansode et.al. presents a comparative analysis of Random forest and Decision Tree and Logistic Regression approach. The algorithm with the highest accuracy is logistic regression, though. Additionally, it can tackle categorical values, scalability issues, overfitting issues, and provide a decent visualization of the data using a confusion matrix. It can also fill in the missing values of the datasets. Poor credit applicants are more likely to be turned down, especially given the possibility of defaulting on the loan. The system can take part in new testing to pass new test cases because it was trained on recent data sets that may get stale in the future.

In [14] Subhiksha et.al. presents a comparative analysis of Logistic Regression and Decision Tree classifier. Starting with data preparation and missing value processing, exploratory analysis is followed by model construction and model evaluation. When we receive greater accuracy scores and other performance indicators, which will be determined, we have the best accuracy on the public test set. This essay can aid in predicting whether a candidate will be granted a bank loan or not.

In [15] Uniyal et.al. presents a fast, easy, and straightforward approach for choosing meritorious applicants. They present a credit risk model built on both homogeneous and heterogeneous classifiers. Three classifiers—the logistic artificial neural network, the logistic regression, and the support vector machine—make up an ensemble model. The findings indicate that when compared to a homogeneous classifiers ensemble, the heterogeneous classifiers ensemble performed better and was more accurate.

In [16] Khan et.al. presents an analysis on the Loan Prediction Models created using different algorithms and select the best one that can speed up loan approval while lowering the risk involved. It is done by making a prediction about whether the loan can be issued to that person based on a number of factors, including credit score, income, age, marital status, gender, etc. The accuracy of the prediction models based on Logistic Regression, Decision Tree, and Random Forest is 80.945%, 93.648%, and 83.388% correspondingly, while the cross-validation results are 80.945%, 72.213%, and 80.130%. This demonstrates that while a decision tree-based model has the highest accuracy for the given dataset, random forest is superior in terms of generalization even though its cross validation is only marginally greater than that of logistic regression.

In [17] Gomathy et.al. presents an assessment and comprehend the process pf prediction on the basis of training on previous datasets. The data model that was produced using a decision tree is used to forecast test results and is then applied to the training data set. There have been instances of mathematical computer errors, content violations, and most importantly, weight of characteristics are fixed in automated prophecy system, so in the near future the so-called software might be made more secure, reliable, and dynamic weight conformation.

In [18] Sunitha et.al. presents the study of prediction that compares the scores, which we got from both the models they have conclude that both the models showing the same and great results on this dataset. Basing on the nature of the output variable they have considering logistic regression as the final model. And the drawback is the accuracy is low when compared with the past models but this model gave an equal importance to both accuracy and precision score. This model became successful in making lower number of False Prediction to reduce the risk factor.

Table 1 Literature Survey

Paper title	**Objective**	**Methodology**	**Inference parameters**	**Pros**	**Cons**
Loan default prediction using decision trees and random forest **[9]**	Individual should be given a loan by assessing certain attributes	Random Forest **[7]**, Decision Tree **[7]**	Correlation heat map between all the features, Box plot of Loan Amount vs Loan purpose	Accuracy of 80% from Random Forest Classifier and accuracy of 73% from Decision Tree approach	Algorithm puts some of the non-defaulters in the default class
Accurate Loan Approval Prediction Based On Machine Learning Approach **[10]**	System approved or rejects the loan applications	Logistic Regression **[7]**, Decision Tree, Random Forest	Loan Train Data set, Loan Train Data set on gender and Dependents	Operationally sound and complies with all Banker specifications	Software could be made more secure, reliable and dynamic weight adjustment
In Banking Prediction of Loan Approbation Using Machine Learning**[11]**	To train the best machine learning model to maximize the predictive capability	Logistic Regression, Decision Tree, Random forest	Accuracy Percentage	Model is slightly superior because it takes into account variables	Software to be improved to make it more dependable, secure, and accurate
Machine Learning Models for Predicting Bank Loan Eligibility **[12]**	Improve the speed, efficacy, and accuracy of loan approval processes	Random Forest, Gradient Boost **[5]**, Decision Tree, Support Vector Machine**[4]**, K-Nearest Neighbour **[5]**, Logistic Regression	Accuracy, Precision , Recall, F1 score	Random forest algorithm scoring the best (95.55%) and the Logistic regression approach scoring the lowest (80%)	Allowing it to participate in fresh testing to pass new test cases

Predicting Loan Approval Using ML **[13]**	Use machine learning concepts to calculate a customer's loan status and predict an immediate and precise outcome	Logistic Regression, Random Forest, SVM, K – Nearest Neighbour, Decision Tree	Confusion Matrix, Accuracy Table, sensitivity, specificity, True Positive, True Negative, False Positive, False Negative	It can categorical values, scalability issues, overfitting issues, and provide a decent visualization of the data using a confusion matrix	Well-defined data set is required for more accurate results
Bank Loan Approval Prediction Using Data Science Technique (ML) **[14]**	Accurately identifies whom to lend loan	Random Forest, Decision Tree, Naive Bayes **[5]**, Logistic Regression	Recall , F1 Score ,Confusion Matrix, Accuracy Table, True Positive, True Negative, False Positive, False Negative	Greater accuracy scores and other performance indicators	Connect with cloud to optimize the work to implement in Artificial Intelligence environment
Loan Approval Prediction using Machine Learning Algorithms Approach **[15]**	Machine learning was used to predict loan acceptance	Logistic Regression, Decision tree, SVM, Random forest	True Positive, True Negative, False Positive, False Negative, Recall , Confusion Matrix, Accuracy	Credit risk model built on both homogeneous and heterogeneous classifiers	Big and well-defined data set is required for more accurate results
Loan Approval Prediction Model: A Comparative Analysis **[16]**	Uses data mining, statistics and probability to forecast an outcome	Prediction model based on Random Forest.	Box plot and histogram are used for study of distribution factors	Logistic Regression, Decision Tree, and Random Forest is 80.945%, 93.648%, and 83.388%	The software to be improved to make it more dependable, secure, and accurate
The Loan Prediction Using Machine Learning **[17]**	Predicting the loan data by using some machine learning algorithms that is Decision Tree	Decision Tree	Recall , Confusion Matrix, Accuracy ,True Positive, True Negative, False Positive, False Negative	Data model that was produced using a decision tree	New testing date should also take part in training data after some fix time
Predicting the Loan Status using Logistic Regression and Binary Tree**[18]**	Extracts the important features of a customer which affects the customer loan status and applies the strategies	Logistic Regression, Binary Tree **[5],**Decision Tree	Accuracy, F1 Score, Precision, Recall	Model gave an equal importance to both accuracy and precision score, model became successful in making lower number of False Prediction to reduce the risk factor	Drawback is the accuracy is low

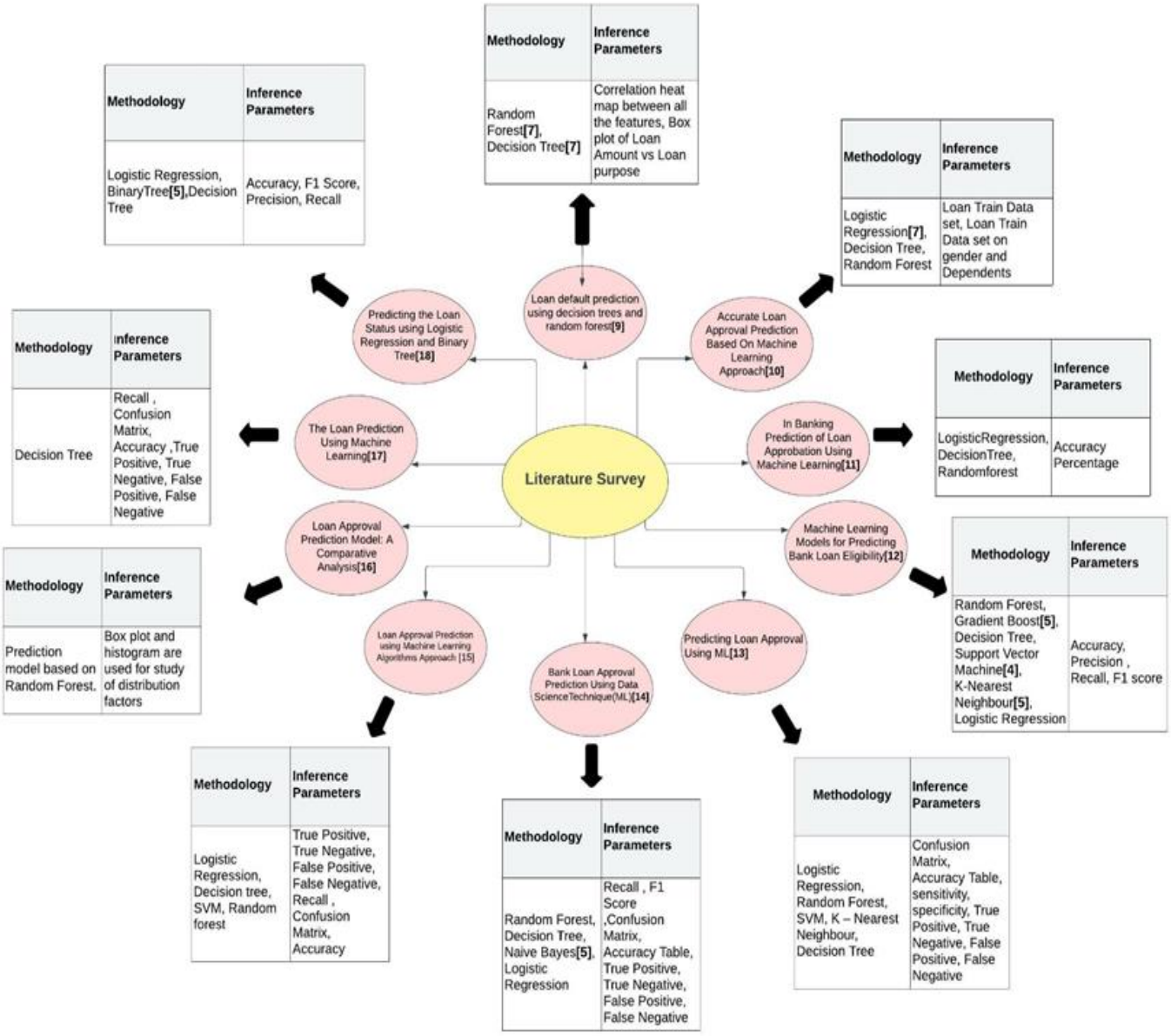

Fig 1 Literature Survey

III. RELATED WORK

In this Section III, we are going to describe the work conducted on "DGHI" dataset for the prediction of customer eligibility in getting an approval of loan from bank whether he is eligible or not. Section III, has been organized into two sub-sections: Section 3.1, explains the regression techniques used for the mentioned objectives namely Logistic Regression, Random Forest and Decision Tree respectively. Section 3.2, explains the description of the dataset used in the research work Section 3.3, explains the methodology of conducting an experimental analysis of the research work conducted in this paper.

3.1. *Description of the Machine Learning Techniques Used:*

In this Section 3.1, we are basically describing the machine learning techniques used in the research work and conducting the experimental analysis of the dataset used in the same. The techniques that have been used in the mentioned work are Logistic Regression, Random Forest and Decision Tree which can be inferred from Figure 2.

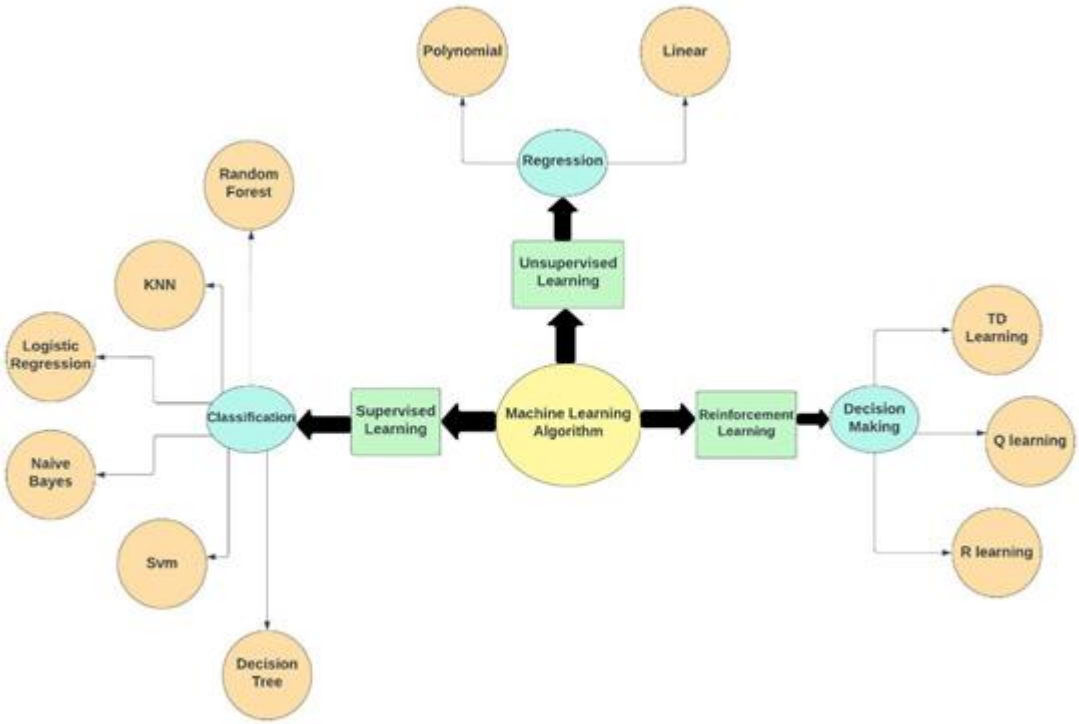

Fig 2 Classification of Machine Learning Techniques

3.1.1. *Logistic Regression:*

One of the most well-known Machine Learning algorithms is logistic regression, which belongs to the Supervised Learning Technique. It is used for predicting the categorical dependent variable using a given set of independent variables. In reality, logistic regression forecasts the results of a categorical dependent variable. As a result, the result must be a discrete or categorical value. Rather of providing the exact values of 0 and 1, it provides the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false and so on.

3.1.1.1. *Features:*

i. The Linear Regression and the Logistic Regression are very similar. Linear Regression is used for solving Regression problems, whereas Classification problems are solved by using Logistic regression.
ii. In logistic regression, we fit a "S" shaped logistic function rather than a regression line, which predicts two maximum values, i.e. (0 or 1).

3.1.1.2. *Advantages:*

i. It makes no assumptions about distributions of classes in feature space.
ii. It is very fast at classifying unknown records.
iii. Good accuracy for many simple datasets and it performs very well when the dataset is linearly separable.

3.1.1.3. *Disadvantages:*

i. If the number of observations is lesser than the number of features, then we have to avoid using logistic regression as it may lead to overfitting.
ii. The assumption of linearity between the dependent variable and the independent variables is the major limitation of Logistic Regression.
iii. It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set.

3.1.1.4. *Applications:*

➢ *Logistic Regression can also be used in the following Areas:*

i. In healthcare, it is used to identify risk factors for diseases and plan preventive measures.
ii. In weather forecasting apps in order to predict snowfall and weather conditions.
iii. In insurance to predict the chances that a policyholder will die before the policy's term expires based on some specific criteria, such as gender, age and physical examination, etc.
iv. In banking to predict the chances that a loan applicant will default on a loan or not, based on annual income, past defaults and past debts.

3.1.2. *Random Forest:*

Random Forest is a popular machine learning algorithm that is a part of the supervised learning methodology. It can be used for both Classification and Regression problems in ML. It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance. Random Forest is a classifier that uses many decision trees on different subsets of the input dataset and averages the results to increase the predictive accuracy of that dataset.

3.1.2.1. *Features:*

i. It needs less time to train than other algorithms.
ii. It operates effectively even for the enormous dataset and predicts output with a high degree of accuracy.
iii. Even when large proportion of data is absent, accuracy can still be maintained.
iv. Random forest's accuracy is typically fairly high.
v. Its efficiency is particularly Notable in Large Datasets.
vi. Provides an estimate of important variables in classification.
vii. Created forests can be saved and used again.
viii. In contrast to other models, it does not over fit with more features.

3.1.2.2. *Advantages:*

i. Both classification and regression tasks can be handled by Random Forest.
ii. It is capable of handling large datasets with high dimensionality.
iii. It improves the model's accuracy and avoids the overfitting problem.

3.1.2.3. *Disadvantages:*

i. As it creates several trees to integrate their outputs, it uses a lot of resources and computational power.
ii. As it integrates numerous decision trees to decide the class, training takes a lot of time.
iii. Due to the ensemble of decision trees, it also suffers from the interpretability and fails to determine the significance of each variable.

3.1.2.4. *Applications:*

i. *Banking:* Banking sector mostly uses this algorithm for the identification of loan risk.
ii. *Medicine:* With the help of this algorithm, disease trends and risks of the disease and many other things can be identified.
iii. *Land Use:* We can identify the areas of similar land use by this algorithm.
iv. *Marketing:* Marketing trends can be identified using this algorithm.

3.1.3. *Decision Tree:*

Decision Tree is a type of Supervised learning technique which is used for the classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, in which the internal nodes represent the features of dataset, branches represent the decision rules and each leaf node represents the outcome.

3.1.3.1. *Features:*

i. It is among the most popular and useful techniques for supervised learning. Decision trees are a non-parametric supervised learning technique used for both classification and regression problems.
ii. The logic behind decision tree can be easily understood because it shows a tree-like structure.
iii. Decision trees are typically designed to resemble how people think when making decisions, and it is easy to understand.

3.1.3.2. *Advantages:*

i. It is easy to use and understand.
ii. Both categorical and numerical data can be handled by the decision tree.
iii. Resistant to the outliers, therefore it requires little data pre-processing.

3.1.3.3. *Disadvantages:*

i. Prone to overfitting.
ii. Require some kind of measurement as to how well they are doing.
iii. Need to be careful with parameter tuning.
iv. Can create biased learned trees if some classes dominate.

3.1.3.4. *Applications:*

i. Marketing:

Businesses can use decision trees to improve the accuracy of their marketing initiatives by keeping track of the performance of their competitors' products and services. Decision trees can help in audience segmentation and support businesses in producing better-targeted advertisements that have higher conversion rates.

ii. Retention of Customers:

Decision trees are used by businesses to retain customers by evaluating their behaviour and developing new offers or goods that correspond to those habits. Companies can determine their clients' levels of satisfaction by employing decision tree models.

iii. Diagnosis of Diseases and Ailments:

Medical personnel can use decision trees to identify individuals who are more likely to develop significant (or preventable) illnesses like diabetes or dementia. Decision trees are very useful since they may reduce options based on particular criteria.

iv. Detection of Frauds:

Companies can prevent fraud by using decision trees to identify fraudulent behaviour beforehand. It can save companies with the lot of resources, including time and money.

3.2. *About the Dataset:*

In this Section 3.2, we describe the dataset that has been used in the research work. Data was gathered from Kaggle **[19]**, one of the most providers of data sources for the purpose of learning, and hence the data is collected from the Kaggle, which had two sets of details, one of which was for the preparation and the supplementary tests. The dataset for training is the model in which datasets are further divided into datasets was used to train the model train and the minor dataset. We will predict the Loan_Status using the model built using the train data. Given below is the description for each variable.

Table 2 About the Dataset

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	Loan approved(Y/N)

3.3. *Proposed Model:*

In this Section, we present the implementation of the proposed model that we have used for conducting the implementation of the proposed work which can be inferred from Figure 3.

- *Step 1. Data Preparation:* In this step, we have converted the raw data into a format (.csv) file format for making the analysis process easy and distinguishable. Followed by process of replacing the missing values with the neighbor's average/mean /median / mode/min./max. Furthermore, primarily for this process we need to import the libraries followed by Step 2.

- *Step 2. Using the Training Set to Fit the Logistic Regression:* In this step, we train our dataset using the random process of inference parameter selection. Then, we apply the LRD. Then using the fit method of sklearn [20] which will train the algorithm on the training data when the model is initiated.

- *Step 3. Foreseeing the Test Outcome:* For this process we follow the process of analysis of the final outcome obtained after Step 2 on the basis of predict () [20]. The main functionality of the predict () in the process of testing of dataset we predict the result on the basis of the data values label using the trained model used in the research work mentioned in Figure 3.

- *Step 4. Evaluate the Result's Correctness:* This involves checking for accuracy. We will check the accuracy of the testing model to that of training model followed by Step 3.

- *Step 5. Displaying the Outcomes of the Test Set:* In this we will visualize the output of the test set in the form of graphical format that has been explained in the Section 4.

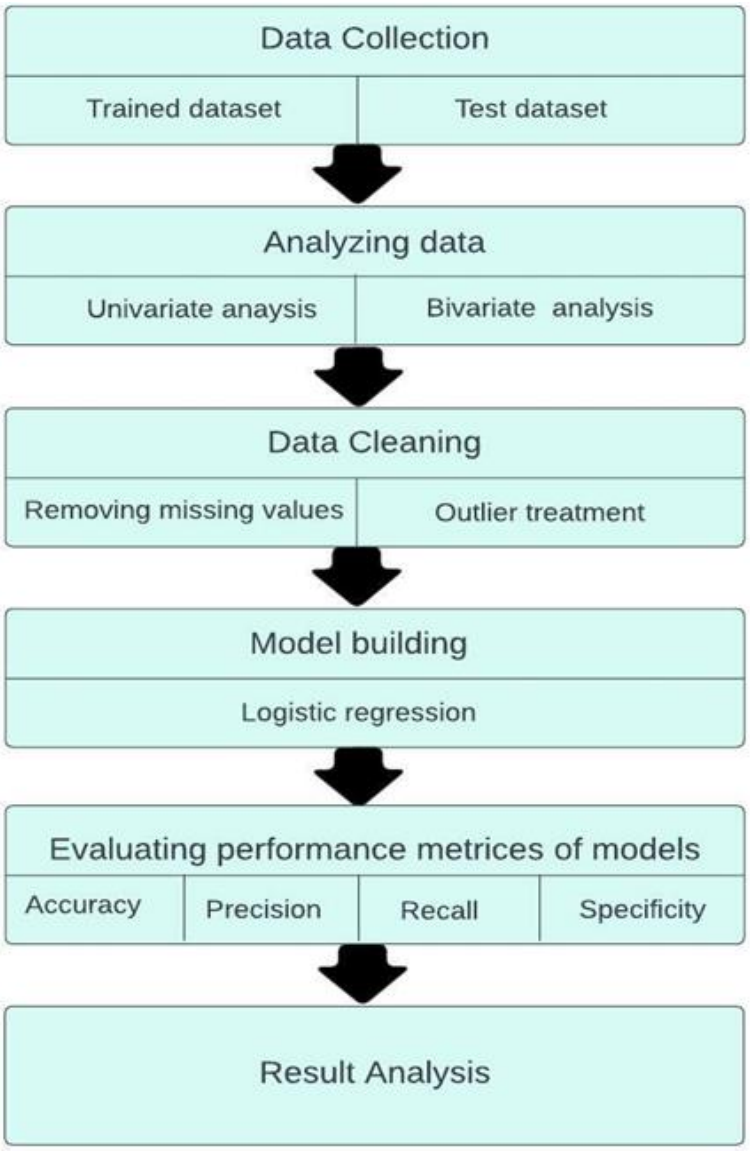

Fig 3 Steps of Implementation

IV. RESULTS AND OBSERVATIONS

In Section 4, we present the analysis and results on the basis of implementation conducted in the Section 3 followed by the methodology represented in the Figure 3. This Section, has been divided into two subsections: Section 4.1, presents the results obtained after conducting the simulation in Python (Jupyter) IDE **[22]**. Section 4.2, gives the observations and conclusion study that we have been obtained from the Section 4.1. Section 4.3, describes the comparative study on the basis of an implementation conducted.

4.1. *Result Analysis:*

In this Section 4.1, we present the simulation study result analysis conducted on the Jupyter using Logistic Regression, Random Forest and Decision Trees respectively on the available dataset [8].

4.1.1. *Using Logistic Regression Model:*

In this Section, we basically analyze the performance of dataset **[8]**. The analysis that has been obtained using the mentioned approach is basically represented X –axis and Y – axis where we denote the False positive value and True positive value respectively. On the basis of the results obtained we analyses that the accuracy rate that has been obtained on the basis of an implementation is $0.7\overline{34}$ which plotted for the ~(900-1000) values of dataset used in our dataset. Accuracy rate denotes the percentage of true predictions obtained after the successful implementation of the dataset. As in our dataset we are basically obtaining the eligibility of the customer whether he is eligible for the loan approval or not from the bank which can be inferred from Figure 4.

```
i=1
kf=StratifiedKFold(n_splits=5,random_state=1,shuffle=True)
for train_index,test_index in kf.split(X,y):
    print('n{} of kfold {}'.format(i,kf.n_splits))
    xtr,xvl = X.iloc[train_index],X.iloc[test_index]
    ytr,yvl = y.iloc[train_index],y.iloc[test_index]
    model = LogisticRegression(random_state=1)
    model.fit(xtr, ytr)
    pred_test = model.predict(xvl)
    score = accuracy_score(yvl,pred_test)
    print('accuracy_score',score)
```

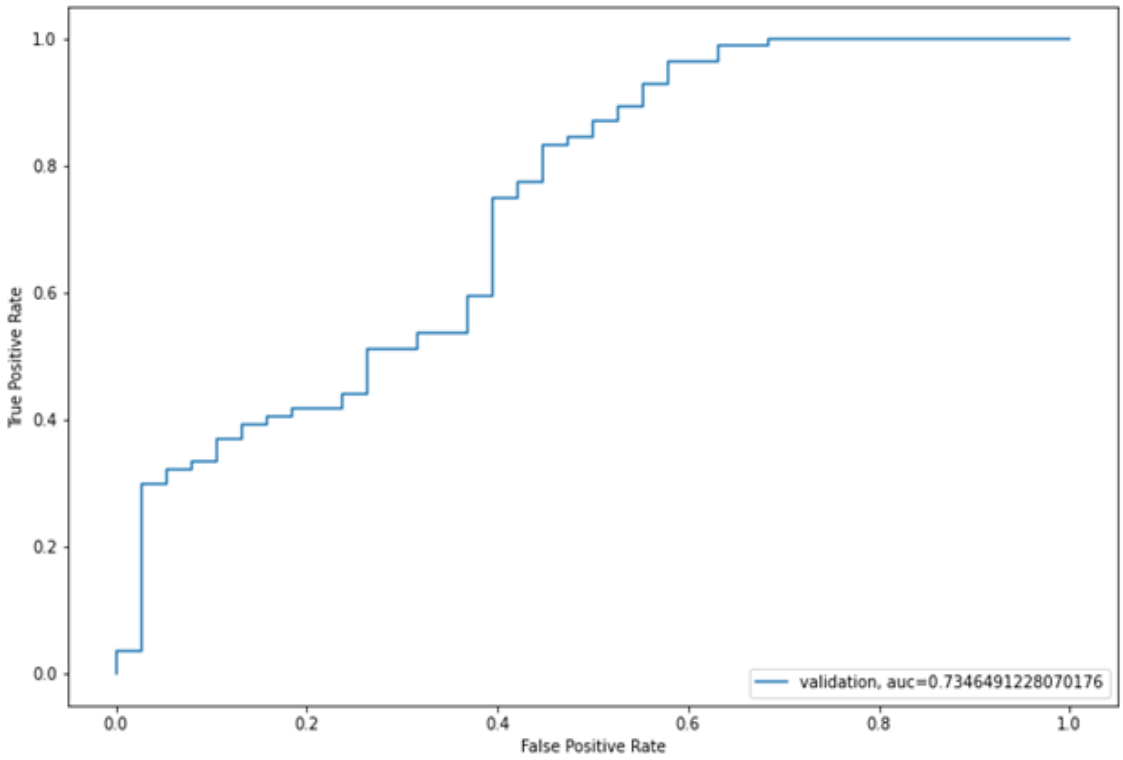

Fig 4 Analysis using Logistic Regression

4.1.2. *Using Random Forest Model:*

Using Random Forest model, we basically analyses the performance of dataset **[8]**. The analysis that has been obtained using the mentioned approach is basically represented X –axis and Y – axis where we denote the False positive value and True positive value respectively. On the basis of the results obtained we analyses that the accuracy rate that has been obtained on the basis of an implementation is $0.6\overline{88}$ which plotted for the ~(900-1000) values of dataset used in our dataset. Accuracy rate denotes the percentage of true predictions obtained after the successful implementation of the dataset. As in our dataset we are basically obtaining the eligibility of the customer whether he is eligible for the loan approval or not from the bank which can be inferred from Figure 5.

```
i=1
kf = StratifiedKFold(n_splits=5,random_state=1,shuffle=True)
for train_index,test_index in kf.split(X,y):
    print('n{} of kfold {}'.format(i,kf.n_splits))
    xtr,xvl = X.loc[train_index],X.loc[test_index]
    ytr,yvl = y[train_index],y[test_index]
    model = RandomForestClassifier(random_state=1, max_depth=10)
    model.fit(xtr, ytr)
    pred_test = model.predict(xvl)
    score = accuracy_score(yvl,pred_test)
    print('accuracy_score',score)
    i+=1
pred_test = model.predict(test)
pred1=model.predict_proba(xvl)[:,1]
```

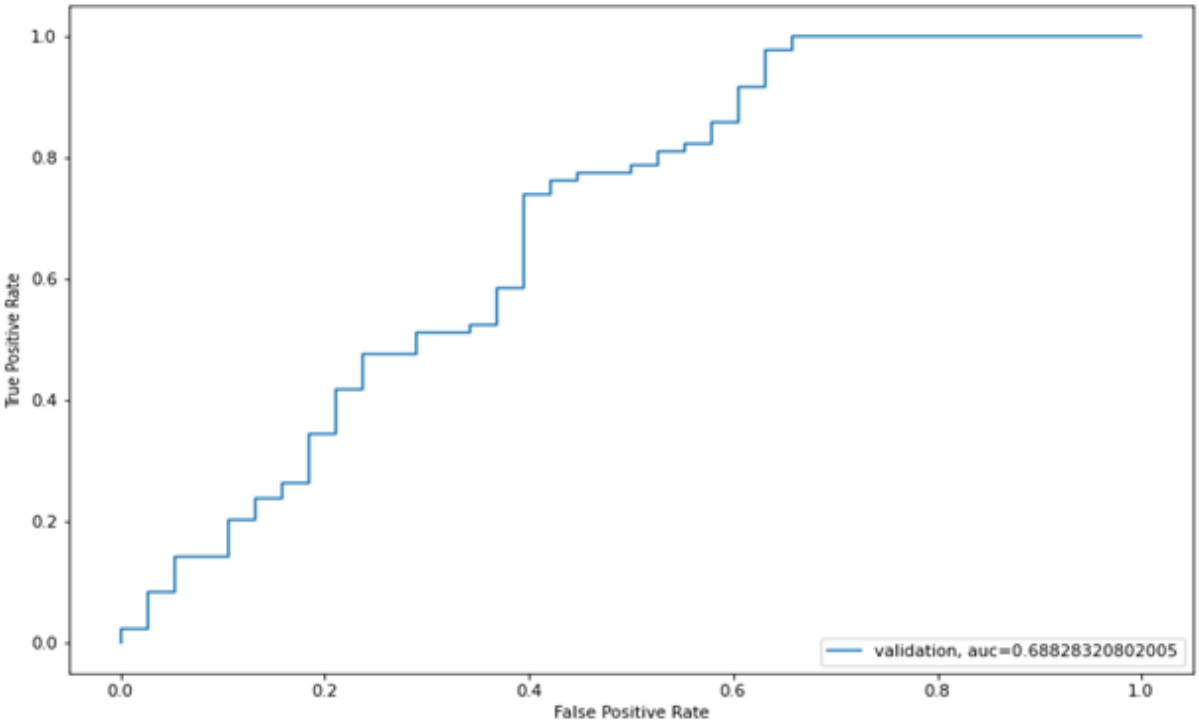

Fig 5 Analysis of Random Forest

4.1.3. *Using Decision Tree Model:*

Using Decision Tree model, we basically analyse the performance of dataset **[8]**. The analysis that has been obtained using the mentioned approach is basically represented X –axis and Y – axis where we denote the False positive value and True positive value respectively. On the basis of the results obtained we analyses that the accuracy rate that has been obtained on the basis of an implementation is $0.6\overline{46}$which plotted for the ~(900-1000) values of dataset used in our dataset. Accuracy rate denotes the percentage of true predictions obtained after the successful implementation of the dataset. As in our dataset we are basically obtaining the eligibility of the customer whether he is eligible for the loan approval or not from the bank which can be inferred from Figure 6.

```
i=1
kf = StratifiedKFold(n_splits=5,random_state=1,shuffle=True)
for train_index,test_index in kf.split(X,y):
    print('n{} of kfold {}'.format(i,kf.n_splits))
    xtr,xvl = X.loc[train_index],X.loc[test_index]
    ytr,yvl = y[train_index],y[test_index]
    model = tree.DecisionTreeClassifier(random_state=1)
    model.fit(xtr, ytr)
    pred_test = model.predict(xvl)
    score = accuracy_score(yvl,pred_test)
    print('accuracy_score',score)
    i+=1
pred_test = model.predict(test)
pred=model.predict_proba(xvl)[:,1]
```

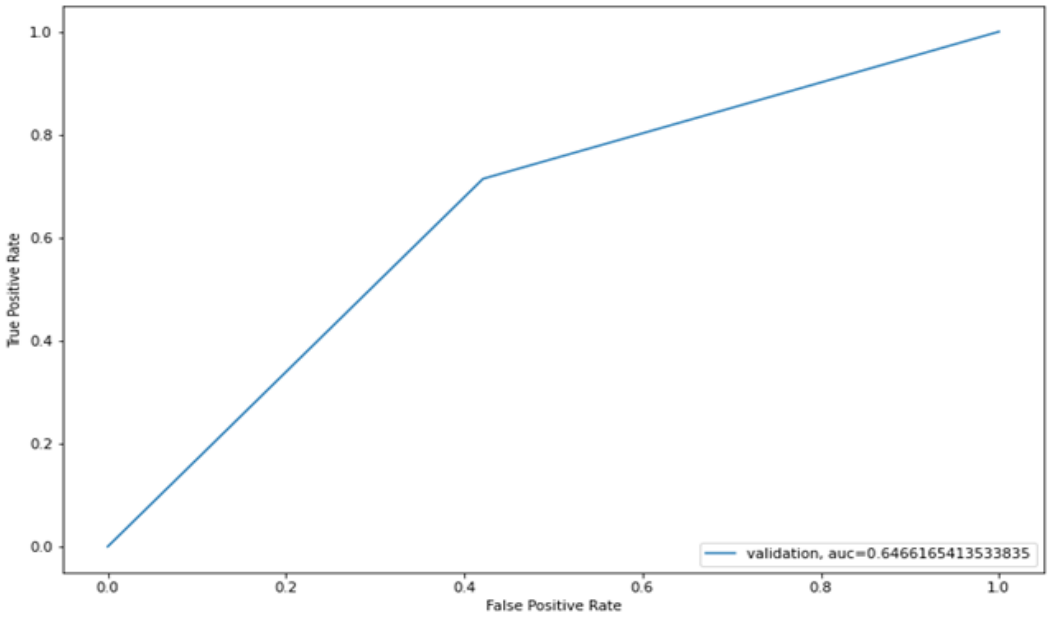

Fig 6 Analysis of Random Forest

4.2. *Observations and Conclusions:*

In this Section, we describe the observations and conclusions on the basis of an implementation done in the former section. The observation study has been done using the Table 3, which also presents the comparative study on the basis of the mentioned features such as: principle, accuracy rate, inference parameters selection, advantages and disadvantages respectively. Also, it presents the conclusion study using the heat map which is generated on the basis of inference parameters namely: Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term, Credit History and Loan Status respectively.

➢ *Heat Map can be Defined in the Following Aspects:*

- Heat map is the graphical representation of the observation wherein the given data values is presented in the form of graphs and colors included all the given parameters chosen with the different colors selection.
- Heat map also defined as “a major responsibility with all the parameters have been represented in the form of graphs with the defined colors along with it the volume of locations/events within a dataset that has been occupied by the variable during the statistical analysis.”
- For the implementation that has been conducted in the given simulation environment chosen has been done using the heat map as shown in Figure 7.

ISSN No:-2456-2165

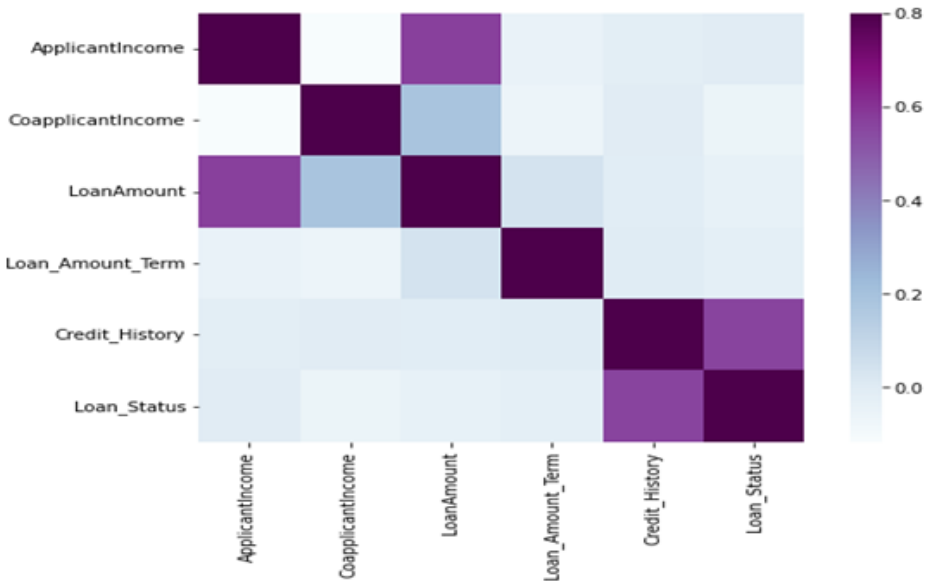

Fig 7 Heat Map

4.3. *Comparative Study:*

On the basis of the simulation and results obtained the comparative study has been presented in the Table 3 that presents the comparative study obtained on the basis of the parameters: principle of the technique used, accuracy rate, inference parameters selection, advantages and disadvantages.

Table 3 Comparative Study of the Results Obtained

S.NO.	COMPARATIVE FEATURES	LOGISTIC	RANDOM FOREST	DECISION TREE
1.	**PRINCIPLE**	It is a machine learning algorithm that is used to predict the probability of certain classes based on some dependent variables	It is a supervised machine learning classification algorithm that is used widely in classification and regression problems.	It is a graphical representation of all possible solutions to a decision based on certain conditions.
2.	**ACCURACY RATE**	73.46%	68.82%	64.66%
3.	**INFERENCE PARAMETERS SELECTION**	1)Total_Income 2)Marital Status 3)Education 4)Credit_History	1)Total_Income 2)Marital Status 3)Education 4)Credit_History	1)Total_Income 2)Marital Status 3)Education 4)Credit_History
4.	**ADVANTAGES**	- It makes no assumptions about distributions of classes in feature space. - It is very fast at classifying unknown records.	- It needs less time to train than other algorithms. - Random forest's accuracy is typically fairly high.	- It is easy to use and understand. - Both categorical and numerical data can be handled by the decision tree
5.	**DISADVANTAGES**	- If the number of observations is lesser than the number of features, then we have to avoid using logistic regression as it may lead to overfitting. - The assumption of linearity between the dependent variable and the independent variables is the major limitation of Logistic Regression.	- As it creates several trees to integrate their outputs, it uses a lot of resources and computational power. - As it integrates numerous decision trees to decide the class, training takes a lot of time.	- Prone to overfitting. - Require some kind of measurement as to how well they are doing.

V. FUTURE WORK

Finally, in our model, we estimate whether or not the loan will be authorized using a logistic regression model. Here, we've created a system for loan acceptance prediction that aids businesses in deciding whether to approve or deny loan requests from clients. The banking sector will benefit from the opening of effective delivery channels. In this model, the prediction is made using the Logistic Regression technique. It is possible to create and evaluate new methods for the domain that perform better than common data mining models. To produce more accurate answers, the model can be trained on enormous amounts of data.

- *Availability of Data and Materials:* The data used to support the findings of this study are included within the article.
- *Competing Interests:* The authors declare that they have no conflict of interest.
- *Funding:* Not Applicable.
- *Author's Contributions:* TG formulated the problem statement and suggested the method of proposed work for resolving an issue of loan prediction in the real – life applications such as banks. The computations and proposed model have been proposed by HS, IT along with an implementation carried out by them. The analysis of proposed work has been measured on the basis of heat map and inference parameters and carried out by GA.
- *Acknowledgments:* We are thankful to IMS Engineering College for providing the research oriented academic environment.

REFERENCES

[1]. Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning." *ACM Computing Surveys (CSUR)* 54, no. 6 (2021): 1-35.

[2]. Caton, Simon, and Christian Haas. "Fairness in machine learning: A survey." *arXiv preprint arXiv:2010.04053* (2020).

[3]. Bahrammirzaee, Arash. "A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems." *Neural Computing and Applications* 19, no. 8 (2010): 1165-1195.

[4]. Abdollahpouri, Himan, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. "Multistakeholder recommendation: Survey and research directions." *User Modeling and User-Adapted Interaction* 30, no. 1 (2020): 127-158.

[5]. Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. "An approach for prediction of loan approval using machine learning algorithm." In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 490-494. IEEE, 2020.

[6]. Kumar, Ashwani, Raman Dugyala, and Pronaya Bhattacharya. "Prediction of Loan Scoring Strategies Using Deep Learning Algorithm for Banking System." In *Innovations in Information and Communication Technologies (IICT-2020)*, pp. 115-121. Springer, Cham, 2021.

[7]. Aslam, Uzair, Hafiz Ilyas Tariq Aziz, Asim Sohail, and Nowshath Kadhar Batcha. "An empirical study on loan default prediction models." *Journal of Computational and Theoretical Nanoscience* 16, no. 8 (2019): 3483-3488.

[8]. https://www.kaggle.com/datasets/shaijudatascience/loan-prediction-practice-av-competition.

[9]. Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain and Preeti Nagrath, 2020. " Loan default prediction using decision trees and random forest: A comparative study". International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020.

[10]. J. Tejaswini, T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni, Venkata Rao Maddumala, 2020. "Accurate Loan Approval Prediction Based On Machine Learning Approach". Journal of Engineering Sciences April/2020.

[11]. Dhanashri A. Wani, Adarsha S. Tajane, Tejal P. Doshi and Sachin B. Bhosale, 2022." In Banking Prediction of Loan Approbation Using Machine Learning". International Journal of Research Publication and Reviews, 31, May 2022

[12]. Ugochukwu.E. Orji, Chikodili.H. Ugwuishiwu, Joseph. C. N. Nguemaleu and Peace. N. Ugwuanyi, 2022." Machine Learning Models for Predicting Bank Loan Eligibility". IEEE NIGERCON, 2022

[13]. Nikhil Bansode, Adarsh Verma, Abhishek Sharma, Varsha Bhole, 2022." Predicting Loan Approval Using ML". International Research Journal of Modernization in Engineering Technology and Science, 05/May-2022

[14]. Subhiksha R, Vaishnavi L, Shalini B, Mr. N. Manikandan, 2022." Bank Loan Approval Prediction Using Data Science Technique (ML)". International Journal for Research in Applied Science & Engineering Technology (IJRASET), 5 May 2022

[15]. Sagar Uniyal, Nitesh Pandey, Ramanand Gupta Vishal Kumar, 2021." Loan Approval Prediction using Machine Learning Algorithms Approach". INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY, June 2021

[16]. Afrah Khan, Eakansh Bhadola, Abhishek Kumar, Nidhi Singh, 2021." Loan Approval Prediction Model: A Comparative Analysis". Advances and Applications in Mathematical Sciences, 2021

[17]. Dr. C K Gomathy, Ms. Charulatha, Mr. AAkash, Ms. Sowjanya, 2021." The Loan Prediction Using Machine Learning". International Research Journal of Engineering and Technology (IRJET), 10 Oct 2021
[18]. T. Sunitha, M. Chandravallik, M. Ranganayaka, G. Suma sri, T.V.S. Jagadeesh, A. Tejaswi, 2020. "Predicting the Loan Status using Logistic Regression and Binary Tree". International Conference on IoT based Control Networks and Intelligent Systems (ICICNIS 2020), 2020.
[19]. ([2016&December]).[Bank Loan Status Data set] . [v er-6].[https ://ww w.k aggle .com /dataset s/zaurbe gi ev/my-dataset]
[20]. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
[21]. Fernando Perez, Brian E Granger, and John D Hunter. Python: an ecosystem for scientific computing. *Computing in Science \\& Engineering*, 13(2):13–21, 2011.