Bank Loan Eligibility Prediction using Machine Learning

Dr.S.N.Tirumala Rao[1], Chilaka Poojitha[2], Gude Sindhu[3], Kamepalli Hima[4]
[1] Professor, [2, 3 & 4] Student
[1]nagatirumalarao@gmail.com, [2]poojithachilaka555@gmail.com, [3] sindhugude1705@gmail.com, [4] kamepallihima90@gmail.com
Department of Computer Science and Engineering,
Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India

ABSTRACT- Machine learning algorithms are transforming operations across various industries, such as finance, real estate, security, and genomics. In the banking sector, one of the most tiresome tasks is the loan approval procedure. Advances in technology, including machine learning models, can enhance the efficiency, precision, and speed of loan approval procedures. This paper presents five (5) machine learning algorithms (Logistic Regression , Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting) for predicting loan eligibility. The models underwent training on the 'Loan Eligible Dataset,' a historical dataset licensed under the Database Contents License (DbCL) v1.0 and accessible on Kaggle. Using Python programming libraries and Kaggle's Jupyter Notebook cloud environment, the dataset was processed and examined. High performance accuracy was demonstrated by our research, with the Random Forest method scoring 96.69%, the highest, and Logistic Regression scoring 76.16%.

KEYWORDS: SVM , SMOTE ,Gradient Boosting techniques, Efficient ML Algorithms , K-Fold , Loan approval prediction.

I. INTRODUCTION

The banking industry is actively seeking to take use of the potential provided by contemporary technologies in order to improve their procedures, boost production, and save expenses, just like many other commercial endeavors.Conventional credit evaluation techniques frequently target underprivileged populations unfairly, hence sustaining socioeconomic inequalities. By implementing machine learning (ML)-driven models, banks may eliminate prejudices and provide more equitable lending access, enabling people and companies from all backgrounds. In 2020, the most sought-after feature for banking apps globally was machine learning's predictive analytics capabilities, according to [1]. The ability of most lending platforms to evaluate credit risk dictates whether they will be successful or not [2]. Any financial organization that grants loans must go through a rigorous procedure. Before extending credit, the bank evaluates a borrower's creditworthiness (defaulter vs. non-defaulter). Creating Machine Learning (ML) models to forecast loan eligibility was the main goal of this study since it helps speed up the decision-making process and determines whether or not an application is approved for a loan[3]. Any knowledge gathered from the past can be used for prediction, and a model is trained to handle incoming input and generate the intended result. The primary advantage of Machine Learning is that it learns from past or previous data and apply that knowledge to predict the outcome. Additionally,ML can enhance the client experience by reducing the need for manual review, offering personalized mortagage recommendations, and making loan decisions more swiftly and accurately. All things considered, machine learning can improve the efficiency and precision of loan processing while simultaneously improving the application experience.

Proposed goals are to: (1) prepare and clean the data for modeling(data preprocessing); (2) analyze the dataset using exploratory data analysis (EDA); (3) construct multiple machine learning (ML) models to forecast loan eligibility; and (4) assess and contrast the various models constructed.

The Machine Learning Algorithms used are Logistic Regression, Support Vector Machine, Decision Tree, Random Forest Classifier and Gradient Boosting.

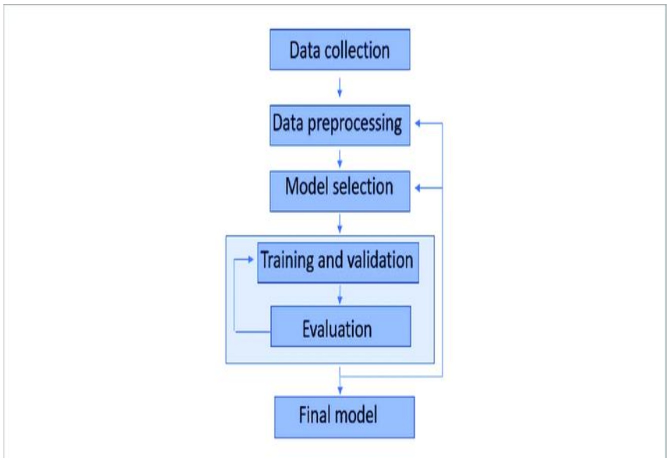


Fig.1 Steps involved in a Model

The sequential phases involved in effective model development are shown in Fig. 1's flowchart. The following stages are covered: gathering data, processing, selecting the model, development and verification, assessment, and final model. To ensure maximum efficiency and effectiveness in creating reliable prediction models, every step of the process has been painstakingly developed to streamline it.

The best-fit models utilizing ML for debt risk estimation were found and contrasted by the authors in [3] after a thorough examination of the literature. The writers wanted to show off the many machine learning algorithms that are employed by scholars to assess the sustainability of rural borrowers, especially those with a bad credit history. Their conclusion demonstrated the broad applicability and excellent performance of the ML algorithms we employed in this study.

II. LITERATURE SURVEY

J. Xu et al.,[4] mentioned how low loan repayment rates negatively affect banks and are a prominent global concern, prompting them to search for more efficient solutions to manage their lending procedures.Utilizing R.F., XGBoost, GBM and ML models based on Neural Networks, the authors evaluated the peer-to-peer (P2P) network in China loan default prediction. All four of their models had accuracy levels above 90%, with RF being the best model. But our objective was to anticipate client loan eligibility, while theirs sought to predict P2P loan default.

H. Meshref et al.,[5] discussed in order to anticipate finance acceptance of information for banking advertising, used a variety of ensemble machine learning approaches in their research, including AdaBoost, LogitBoost, Bagging, and RF model. According to their study, AdaBoost had the best accuracy, scoring 83.97%.

As stated by A.S. Aphale et al.,[6], the study used real the financial institution credit records to forecast consumers' employability and assist banks in developing a digital threat evaluation tool. Several machine learning methods were employed, including KNN, naive Bayes, decision trees, neural networks, ensemble learning, and neural networks. From 80% to 76%, respectively, was the range of their model accuracy.

Partheeban.G et al., [7], examined several machine learning techniques in order to project funding approval outcomes with precision. The process of building the predictive model is covered, including gathering data, preprocessing, choosing a model, training, and validating it. The aim of this project is to boost the effectiveness and precision of loan approval systems by utilizing innovative methods of ML. All in all, it offers a thorough strategy for utilizing cutting-edge data analytics methods to streamline the loan approval procedure.

Tripti et al.,[8] suggested that it revolves around a classification problem, which is a type of supervised learning where it's crucial to ascertain whether the loan will be granted or denied. Additionally, it is a predictive modeling issue in which, for a given sample of input data, a class label is predicted from the input data. They used a variety of ML approches in their work to contrast the effectiveness of implemented prototypes and determine the state of loan approval. Using data from the loan eligibility prediction dataset acquired from Kaggle to predict the target column on the test dataset. Measured characteristics include accuracy, precision, ROC curve, and confusion matrix.

Rabiatus et al., [9] explained that Random Forest algorithm (RF) leverages a recursive binary splitting technique to arrive at the ultimate nodes within a classification and regression tree structure.The comprehensive overview of this literature summarizes the body of knowledge regarding machine learning-based creditworthiness prediction, the authors examine the techniques, datasets, and performance indicators used in pertinent studies. The evaluation provides insights into the advantages and disadvantages of various methodologies and makes recommendations for future lines of inquiry.

Sangeetha et al.,[10] revealed that machine learning (ML) techniques are used to forecast loan-deserving candidates by extracting patterns from a dataset of applicants who were granted for loans.. The customer's prior information will be utilized, such as their work status, income, and most recent credit report. To find the most relevant features, few machine learning techniques are employed. 82% was attained by Random Forest. When compared to other algorithms, the Random Forest algorithm performs better.

These papers show that many strategies and techniques were used in machine learning-based bank loan eligibility prediction. Scholars persistently investigate novel approaches to improve credit evaluation procedures and reduce hazards in lending activities, ranging from conventional algorithms to state-of-the-art deep learning.

III. PROPOSED SYSTEM

The following criteria serve as the foundation for our model's proposal

Dataset Analysis

Data Visualization

Preprocessing Techniques

Model Creation and Evaluation

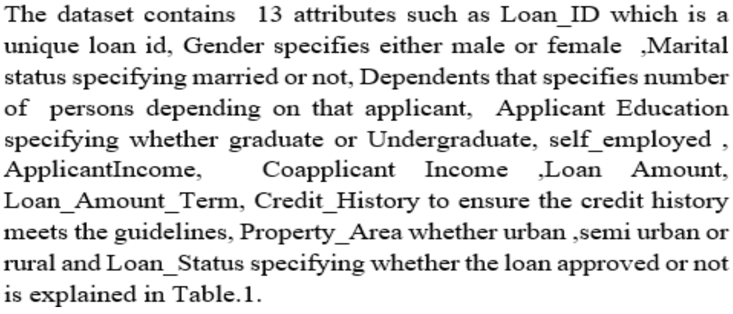
Accuracy

A. Dataset Analysis

The dataset is downloaded from the Kaggle website via the internet. Python was used for this study on the cloud Jupyter Notebook environment provided by Kaggle[11]. In order to perform any kind of prediction, we certainly require earlier bank loan datasets. We have gathered the LoanApprovalPrediction.csv dataset.

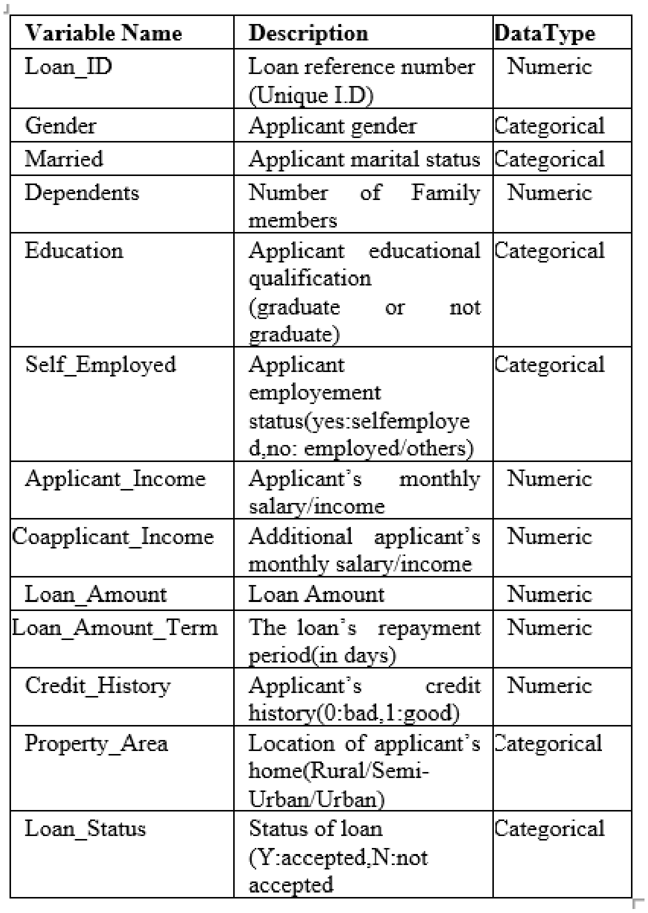
Using the facts at hand, the suggested model predicts a customer's eligibility for a loan. As displayed in Fig.2, the model's input consists of attributes from the dataset. The model's output indicates whether or not the customer qualifies for the loan. The dataset is covered in the next section along with an explanation of the techniques used to clean and prepare the dataset for modeling.

Dataset:



The dataset contains 13 attributes such as Loan_ID which is a unique loan id, Gender specifies either male or female ,Marital status specifying married or not, Dependents that specifies number of persons depending on that applicant, Applicant Education specifying whether graduate or Undergraduate, self_employed , ApplicantIncome, Coapplicant Income ,Loan Amount, Loan_Amount_Term, Credit_History to ensure the credit history meets the guidelines, Property_Area whether urban ,semi urban or rural and Loan_Status specifying whether the loan approved or not is explained in Table.1.

Table.1 LoanApprovalPrediction.csv



Variable Name	Description	DataType
Loan_ID	Loan reference number (Unique I.D)	Numeric
Gender	Applicant gender	Categorical
Married	Applicant marital status	Categorical
Dependents	Number of Family members	Numeric
Education	Applicant educational qualification (graduate or not graduate)	Categorical
Self_Employed	Applicant employement status(yes:selfemployed,no: employed/others)	Categorical
Applicant_Income	Applicant's monthly salary/income	Numeric
Coapplicant_Income	Additional applicant's monthly salary/income	Numeric
Loan_Amount	Loan Amount	Numeric
Loan_Amount_Term	The loan's repayment period(in days)	Numeric
Credit_History	Applicant's credit history(0:bad,1:good)	Numeric
Property_Area	Location of applicant's home(Rural/Semi-Urban/Urban)	Categorical
Loan_Status	Status of loan (Y:accepted,N:not accepted	Categorical

B. Data Visualization

A Graphical representation which is visual representation of dataset is made in order to get brief understanding and visualization of the attributes that are required to predict the loan status.

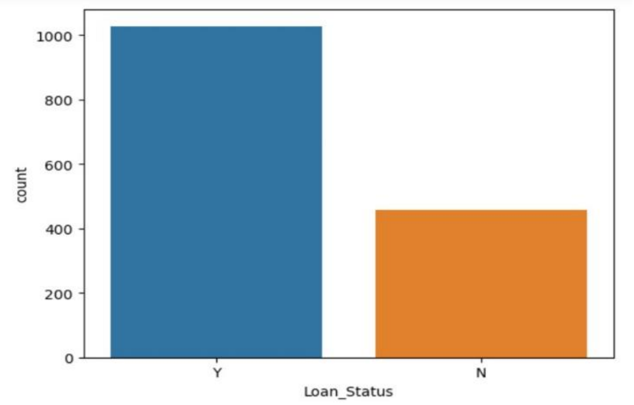


Fig.2 Loan_Status of Applicant

The Fig.2 shows the visual representation of count of Loan_Status of different Applicants. The count of Loan_Status is recorded as Yes and No.

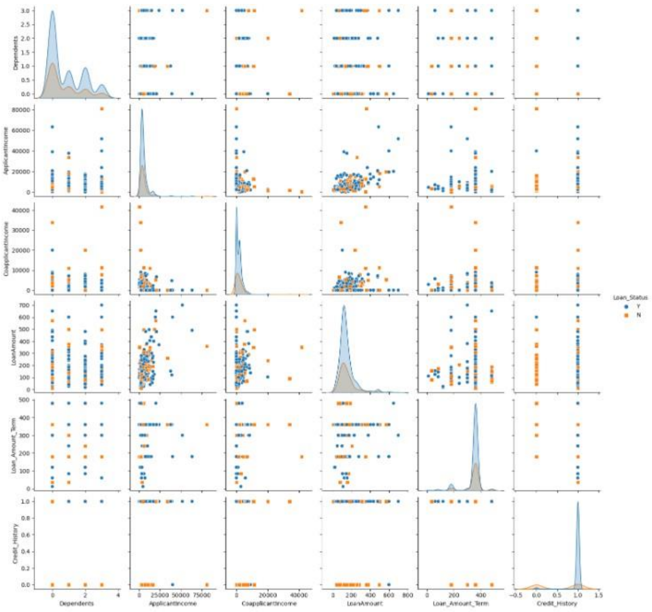


Fig.3 Pair plot of Attributes

The Fig.3 shows the Pair Plot of different Attributes that are required for Loan Approval.
In this Pair Plot the relation between each pair of variables and also the distribution of individual variables is visualized . It provides a distinctive summary of the distributions and correlations within the dataset by combining scatter plots and histograms.

C. Pre Processing Techniques

Before the data is given to the algorithm, it undergoes a procedure known as pre-processing. The large number of occurrences in the dataset means that there will be noise in the data, which may affect how well the model performs.
To guarantee the model to operate at its best, the data was assessed and prepared using the following techniques:

1. Handling null values: Initial data preprocessing involves addressing duplicates and missing values (nulls). Because null values might introduce biases, machine learning models may behave differently when they are present in the dataset. In order to remove this unnecessary data, Pre-processing includes actions such as substituting mean, mode, and median for null values that are present in existing data. Similar to null values, noise can also occasionally be substituted with instances of null values.

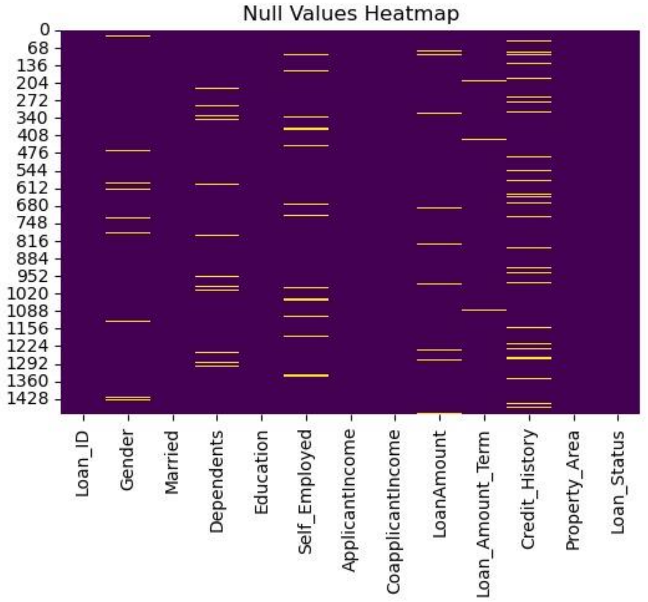


Fig. 4. Heatmap before removing null values

In the Fig.4 the Heatmap which shows the null values in the dataset is visualized.

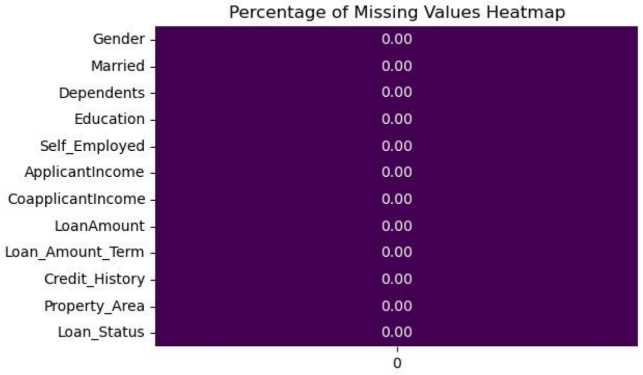
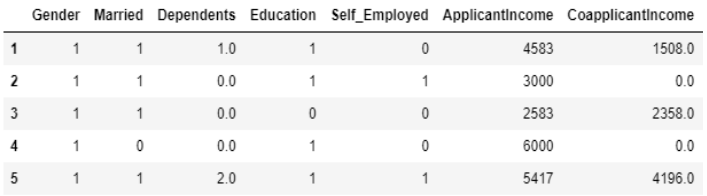


Fig. 5. Heatmap after removing null values

2. TheSyntheticMinorityOversamplingTechnique(SMOTE): It is a useful strategy for addressing imbalanced classification issues, which are a major cause of mistake in machine learning models. The imbalance happens when the minority class has a small proportion of the dataset, which makes it difficult for a model to efficiently discover the decision-making limit[12].
   In the proposed work, oversampling the minority class cases is done using the SMOTE technique to get around this problem. The training dataset is used to create duplicates of the minority class prior to model fitting.

3. One Hot Encoding: The one-hot encoding technique helps transform categorical attributes in a dataset into binary form in order for the ML model to comprehend the data. Additionally, column transformer, which separates the numerical and categorical data, works better with numerical data which can be shown in the fig.7.

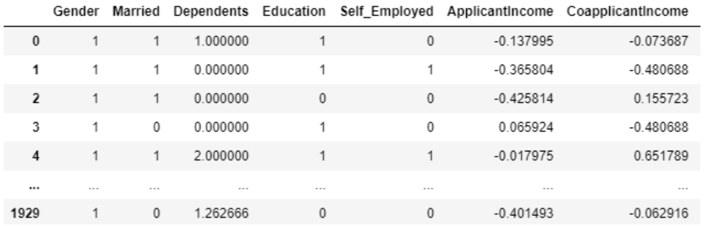


	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome
1	1	1	1.0	1	0	4583	1508.0
2	1	1	0.0	1	1	3000	0.0
3	1	1	0.0	0	0	2583	2358.0
4	1	0	0.0	1	0	6000	0.0
5	1	1	2.0	1	1	5417	4196.0

Fig. 6. Encoding

The Fig.6 gives the data after converting the the categorical variables like gender, married, self_employed and so on into numerical data which is either 1 or 0.

4. Normalization: Transforming features and ensuring that they are all on the same scale is the aim of normalizing data for machine learning models. Normalization improves the model's performance and training stability.In this research , the normalization techniques that are used are standard scalar.



	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome
0	1	1	1.000000	1	0	-0.137995	-0.073687
1	1	1	0.000000	1	1	-0.365804	-0.480688
2	1	1	0.000000	0	0	-0.425814	0.155723
3	1	0	0.000000	1	0	0.065924	-0.480688
4	1	1	2.000000	1	1	-0.017975	0.651789
...	...	...	...	...	...	...	...
1929	1	0	1.262666	0	0	-0.401493	-0.062916

Fig. 7. Normalization

The Fig.7 depicts the data after applying the standard scaler technique to the dataset.Standard Scalar adjusts the information to have a mean (μ) of 0 and a standard deviation (σ) of 1.

5. Exploratory Data Analysis (EDA) is a method that entails looking through the dataset to find abnormalities, trends, and patterns. At this point, the dataset was additionally cleaned to eliminate or manage incomplete or missing data using data imputation, which involves replacing missing values with approximate estimates.
   After looking over the dataset, it is observed that:
   - There are more male applications than female applicants.
   - There are more married applicants in the dataset.
   - There are more applications with good credit (1) in the dataset than there are applicants with negative credit (0).

   To clearly understand how the attributes are related to one another, refer to the Correlation Matrix of attributes shown in the above Fig.8 . One statistical method for assessing the connection of a pair of variables in a set of data is to create a correlation matrix. The matrix is a table in which each cell's correlation value (0 being neutral, -1 being weak), and 1 representing the relationship value for every cell, to show how strongly the variables are associated.

   It is evident from the Fig.8 that the two attributes that are most closely related to each other are LoanAmount and ApplicantIncome with the percentage of 0.52.

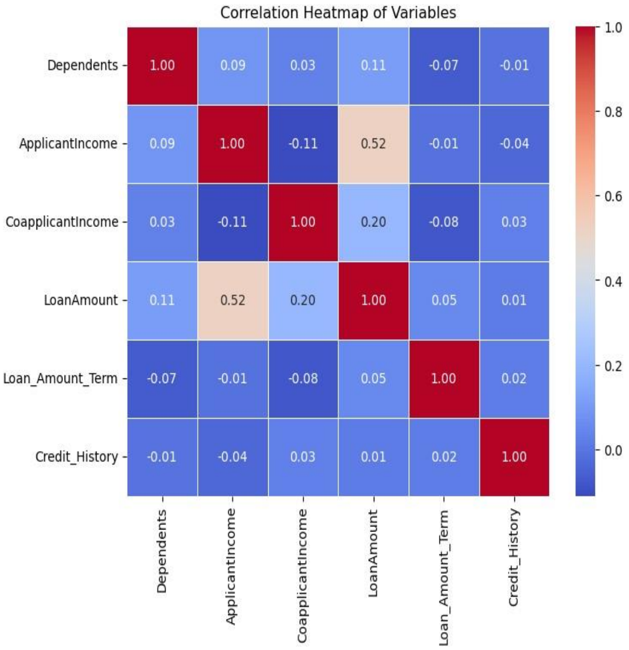


Fig. 8. Correlation Matrix

Ultimately, the quality and performance of the models can be greatly impacted by the data preprocessing stage, which is critical to data analysis and machine learning.

Properly preprocessing the data can help avoid issues such as bias, over fitting, and poor model performance.

D. Creation and Evaluation of Model

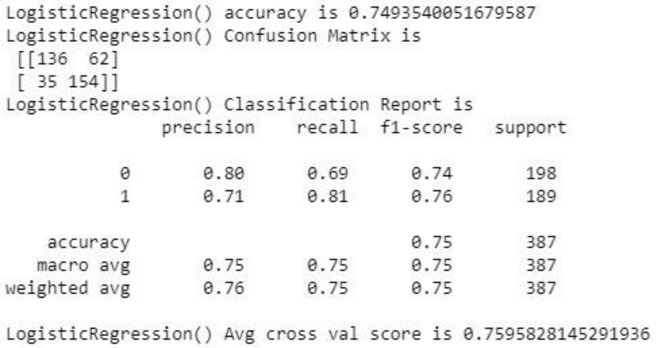
Selecting a suitable ML technique and then training it on the available data are the tasks of this step. By decreasing the discrepancy between the model's predicted output and the actual output found in the training set, the parameters of the model are optimized. The model is tested on a different validation dataset to determine its performance after training. The kind of problem and the performance standards determine the assessment measures that are employed. By modifying its settings or choosing an alternative procedure, the model can be further improved in light of the evaluation findings. This is a crucial step in enhancing the model's performance using fresh, untested data. The model can be used to make predictions or choices on fresh data in a production setting after it has been trained and assessed. To create and evaluate a machine learning model, there are a number of phases involved that need to be carefully thought out and meticulously attended to. The processes that are able to learn to learn from fresh information and generate precise estimation can be developed by following these steps.

The F1 Score and Confusion Matrix are two performance indicators that were employed in the proposed work.The confusion matrix of an ML model classifies and summarises the number of correct and incorrect predictions the model made.

a. "True positive"(TP) :actual positive cases that are correctly predicted by the model.
b. "False positive"(FP) :actual positive cases that the model predicts incorrectly.
c. "True negatives"(TN) :actual negative cases that the model predicts correctly.
d. "False negative"(FN) :actual negative cases that the model predicts incorrectly.
e. Accuracy : Accuracy, which is defined as the ratio of correctly predicted observations to all observations, is the most important performance statistic. Higher accuracy levels are indicative of more accurate responses from the model; however, they are only achievable with symmetric datasets, which have almost equal false positive and false negative values.
   Accuracy = (TP+TN) / (FP+FN+ TN+TP)[13]

i. LogisticRegression(LR)Algorithm:

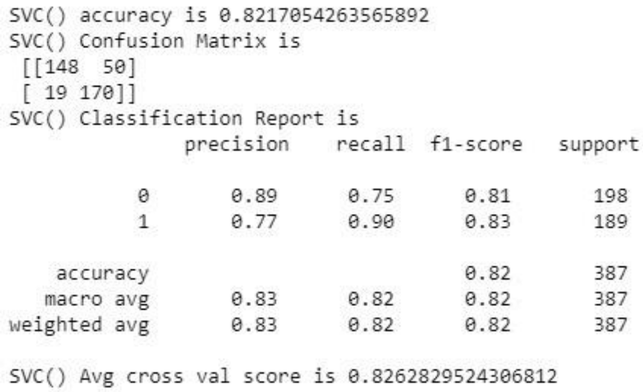
A machine learning classification supervised procedure called logistic regression[14] is used to forecast the likelihood of a categorical dependent variable. It is a statistical technique for analyzing datasets in which an outcome is determined by one or more independent factors. Finding the appropriate model to describe the relationship between a set of independent factors and dependent variables is the main goal of logistic regression. The dependent variable in logistic regression is binary and comprises data that is coded as either 0 (no, failure, etc.) or 1 (yes, success, etc.).[15].In the proposed model the Logistic Regression algorithm achieved the accuracy of 74.93 and after applying the k-fold the accuracy increased to 75.95.



```
LogisticRegression() accuracy is 0.7493540051679587
LogisticRegression() Confusion Matrix is
 [[136  62]
 [ 35 154]]
LogisticRegression() Classification Report is
              precision    recall  f1-score   support

           0       0.80      0.69      0.74       198
           1       0.71      0.81      0.76       189

    accuracy                           0.75       387
   macro avg       0.75      0.75      0.75       387
weighted avg       0.76      0.75      0.75       387

LogisticRegression() Avg cross val score is 0.7595828145291936
```

ii. Support Vector Machine(SVM):

The SVM algorithm specifically looks for a hyperplane in a set of N characteristics that may be used to categorize data points (vectors) in a dataset in a unique way [16].

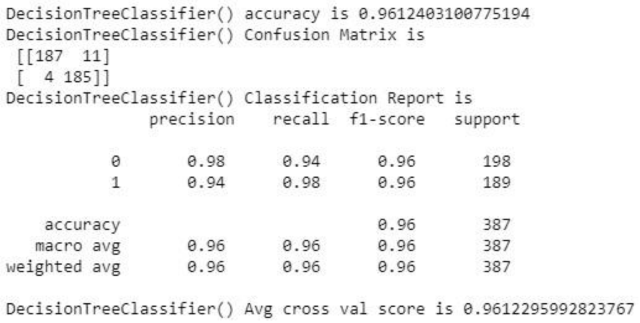


```
SVC() accuracy is 0.8217054263565892
SVC() Confusion Matrix is
 [[148  50]
 [ 19 170]]
SVC() Classification Report is
              precision    recall  f1-score   support

           0       0.89      0.75      0.81       198
           1       0.77      0.90      0.83       189

    accuracy                           0.82       387
   macro avg       0.83      0.82      0.82       387
weighted avg       0.83      0.82      0.82       387

SVC() Avg cross val score is 0.8262829524306812
```

Additionally, the SVM method is renowned for its improved speed and efficiency when working with a small number of data samples[17].In the proposed model the Support Vector Machine achieved an accuracy of 82.62 before and after applying the k-fold.

iii. Decision Tree (DT) Algorithm:

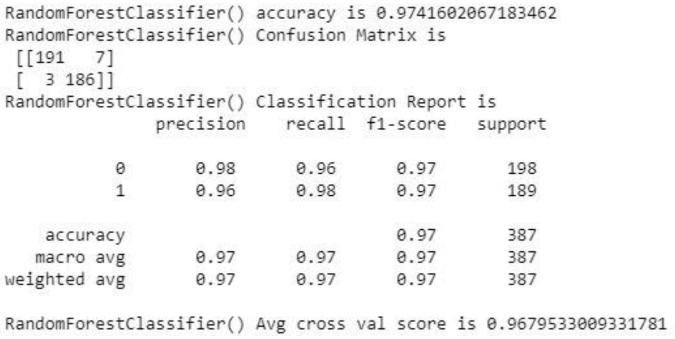
Using a decision tree, one may construct classification models that resemble trees. Essentially, it creates a decision tree for each of the progressively smaller subgroups that are created from the dataset. Every branch of a decision node, which has two or more of them, leads to either another decision node or a leaf node, which represents the final choice[18]. Decision trees are employed in the management of numerical and category data. It classifies datasets using an exhaustive and mutually exclusive collection of if-then rules. One by one, the rules are being learned using the training dataset in a sequential manner. Every time a rule is learned, its associated tuples are removed. This procedure is followed on the training set until a termination condition is met.It is built using a divide-and-conquer recursive top-down approach, and every attribute in the dataset needs to be categorical. Alternatively, they ought to be pre-discretized. The characteristics at the top of the tree have a bigger impact on classification when using the information gain strategy.[19].In the proposed work the Decision tree algorithm achived the accuracy of 96.12 before and after applying of k-fold.



```
DecisionTreeClassifier() accuracy is 0.9612403100775194
DecisionTreeClassifier() Confusion Matrix is
 [[187  11]
 [  4 185]]
DecisionTreeClassifier() Classification Report is
               precision    recall  f1-score   support

           0       0.98      0.94      0.96       198
           1       0.94      0.98      0.96       189

    accuracy                           0.96       387
   macro avg       0.96      0.96      0.96       387
weighted avg       0.96      0.96      0.96       387

DecisionTreeClassifier() Avg cross val score is 0.9612295992823767
```

iv. Random Forest (RF) Algorithm:

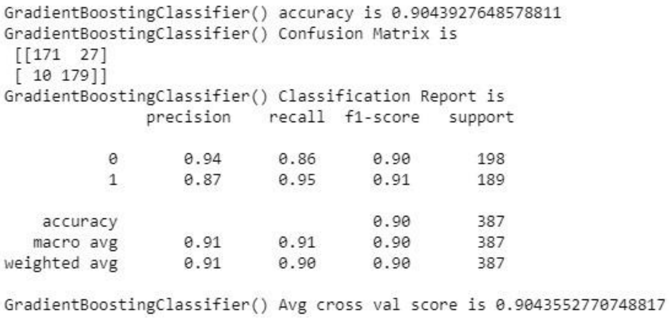
Random Forest is a supervised ML algorithm that makes use of the collaborative learning strategy. Through the use of ensemble learning, many algorithm types can be combined or apply the same algorithm repeatedly to create extremely potent prediction models. It creates a lot of decision trees during training and reports the class, or the average of the classes of each individual tree. Random decision forests counteract decision trees' propensity for incorrectly fitting their training set.[20].In the proposed work the Random Forest algorithm attained the precision of 96.79 before and after applying the k-fold technique.



```
RandomForestClassifier() accuracy is 0.9741602067183462
RandomForestClassifier() Confusion Matrix is
 [[191   7]
 [  3 186]]
RandomForestClassifier() Classification Report is
               precision    recall  f1-score   support

           0       0.98      0.96      0.97       198
           1       0.96      0.98      0.97       189

    accuracy                           0.97       387
   macro avg       0.97      0.97      0.97       387
weighted avg       0.97      0.97      0.97       387

RandomForestClassifier() Avg cross val score is 0.9679533009331781
```

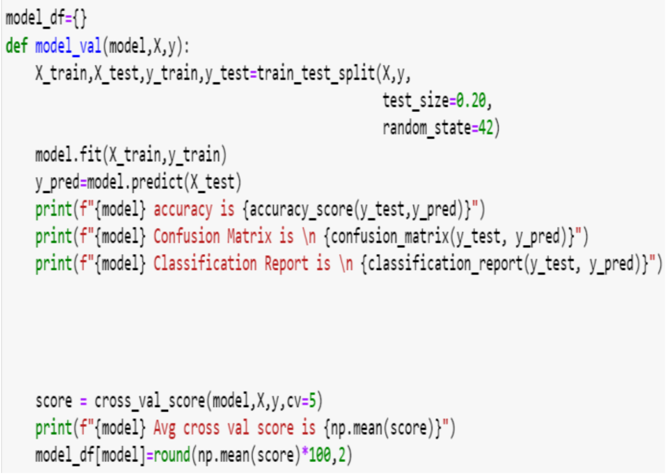
v. Gradient Boosting (GB) Algorithm:

The boosting method does, however, have an impact on a special class of algorithms tasked with turning weak learners into strong learners. This is achieved by figuring out how accurate the weak classifiers are, iteratively learning from them, and merging them into a final robust classifier. [21]. The boosting approaches reweight the training set and assign weights to any misclassified occurrences in the sequence after each iteration. [22]. The Gradient Boost (GBM) Algorithm is the boosting technique used in the proposed work.It achieved the accuracy of 90.43.

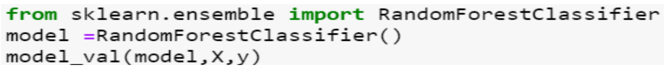


```
GradientBoostingClassifier() accuracy is 0.9043927648578811
GradientBoostingClassifier() Confusion Matrix is
 [[171  27]
 [ 10 179]]
GradientBoostingClassifier() Classification Report is
               precision    recall  f1-score   support

           0       0.94      0.86      0.90       198
           1       0.87      0.95      0.91       189

    accuracy                           0.90       387
   macro avg       0.91      0.91      0.90       387
weighted avg       0.91      0.90      0.90       387

GradientBoostingClassifier() Avg cross val score is 0.9043552770748817
```

Example code:

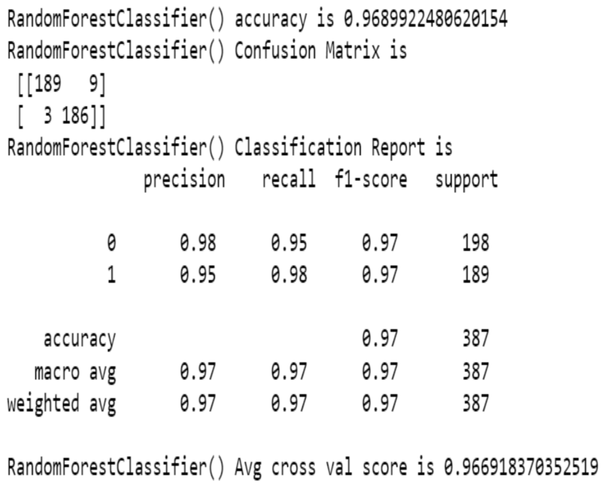


```
model_df={}
def model_val(model,X,y):
    X_train,X_test,y_train,y_test=train_test_split(X,y,
                                                   test_size=0.20,
                                                   random_state=42)
    model.fit(X_train,y_train)
    y_pred=model.predict(X_test)
    print(f"{model} accuracy is {accuracy_score(y_test,y_pred)}")
    print(f"{model} Confusion Matrix is \n {confusion_matrix(y_test, y_pred)}")
    print(f"{model} Classification Report is \n {classification_report(y_test, y_pred)}")




    score = cross_val_score(model,X,y,cv=5)
    print(f"{model} Avg cross val score is {np.mean(score)}")
    model_df[model]=round(np.mean(score)*100,2)
```



```
from sklearn.ensemble import RandomForestClassifier
model =RandomForestClassifier()
model_val(model,X,y)
```

The output of the following code will be as following:



```
RandomForestClassifier() accuracy is 0.9689922480620154
RandomForestClassifier() Confusion Matrix is
 [[189   9]
 [  3 186]]
RandomForestClassifier() Classification Report is
              precision    recall  f1-score   support

           0       0.98      0.95      0.97       198
           1       0.95      0.98      0.97       189

    accuracy                           0.97       387
   macro avg       0.97      0.97      0.97       387
weighted avg       0.97      0.97      0.97       387

RandomForestClassifier() Avg cross val score is 0.966918370352519
```

In the same way the given example code is applied to remaining algorithms that are used in the proposed work.

IV. Results and Analysis

Accuracy is one often used measure to evaluate the performance of a ML algorithm. It calculates the percentage of each instance in the test dataset that is correctly classified.In Predicting Bank Loan Eligibility we used five models of ML such as LR, SVM, DT, RF and GB Algorithm.

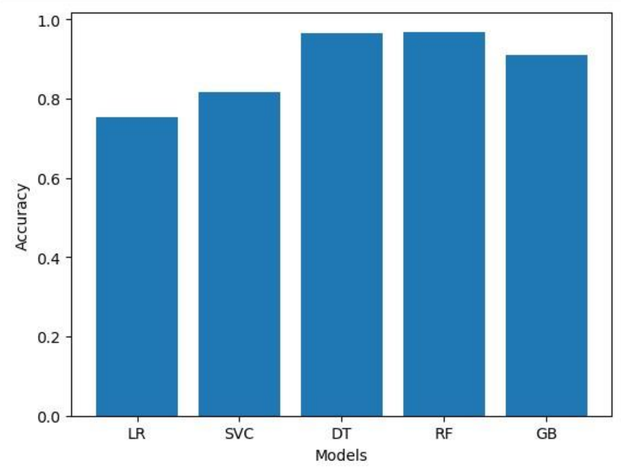


Fig. 9. Accuracy scores of different model.

The Fig.9 shows the visual representation of accuracies of different models that are used in the proposed work.

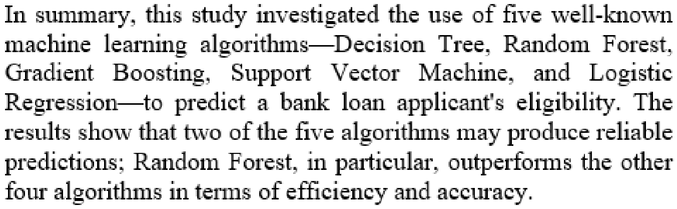
Table-2: Accuracy of ML models

ALGORITHM	EXISTED	PROPOSED
LogisticRegression	80%	76.16%
Support Vector Machine	84.4%	82.67%
Decision Tree	91.1%	95.65%
Random Forest	95.56%	96.79%
Gradient Boosting	93.3%	90.43%

In the Table.2 it is clearly given, the accuraries of each and every model that is used in our research compared with the accuracies in the existed model. In the existed model the Logistic regression model achieved an accuracy of 80%,Support Vector Machine of 84.44%,Decision Tree of 91.11%,Random Forest of 95.56% and Gradient Boosting of 93.33% accuracy. In the proposed model the Logistic Regression achieved the accuracy of 76.16%,Support Vector Machine of 82.67%,Decision Tree of 95.65%,Random Forest of 96.79% and Gradient Boosting of accuracy 90.43%.

V. CONCLUSION AND FUTURE SCOPE

Failure to pay the loan is a major economical risk to the financing sector since it can have a negative impact on lenders' interests and erode public trust. The academic community has committed considerable resources to the development of operational methods of ML in order to support authorities in executing a precise authorization for the loan in real-time.



In summary, this study investigated the use of five well-known machine learning algorithms—Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, and Logistic Regression—to predict a bank loan applicant's eligibility. The results show that two of the five algorithms may produce reliable predictions; Random Forest, in particular, outperforms the other four algorithms in terms of efficiency and accuracy.

In our proposed model the accuracy is increased by applying K-fold technique to the models.

Our model performed quite accurately as measured by the metrics for recall as well as accuracy; the R.F. model scored 96%.For the purpose of forecasting a loan applicant's eligibility for a bank loan, banks and financial aid providers may find it helpful to utilize the predictive model that this study generated. Overall, the outcomes of this experiment show how ML may raise the accuracy and consistency of loan eligibility estimates in the banking industry. Subsequent investigations can build upon this.

REFERENCES

[1] Most commonly used A.I. application in investment banking worldwide 2020, by types." Statista, 15-Sept-2021 [Online]. Available:https://www.statista.com/statistics/1246874/ai-used-in-investment-bankingworldwide-2020/ [Accessed: 29-Jan-2022]

[2] Dorfleitner, G., Oswald, EM. & Zhang, R. From Credit Risk to Social Impact: On the Funding Determinants in Interest-Free Peer-to-Peer Lending. *J Bus Ethics* **170**, 375–400 (2021). https://doi.org/10.1007/s10551-019-04311-8

[3] Kumar, A.; Sharma, S.; Mahdavi, M. Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review. *Risks* **2021**, *9*, 192. https://doi.org/10.3390/risks9110192

[4] Xu, J., Lu, Z. & Xie, Y. Loan default prediction of Chinese P2P market: a machine learning methodology. *Sci Rep* **11**, 18759 (2021). https://doi.org/10.1038/s41598-021-98361-6

[5] Meshref, H. (2020). Predicting loan approval of bank direct marketing data using ensemble machine learning algorithms. *International Journal of Circuits, Systems and Signal Processing*, *14*, 914-922. https://doi.org/10.46300/9106.2020.14.117

[6] A.S. Aphale, and S.R. Shinde, 2020 "Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval." International Journal of Engineering Research & Technology (IJERT)., Vol. 9 pp. 991-995

[7] V. Singh, A. Yadav, R. Awasthi and G. N. Partheeban, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach," *2021 International Conference on Intelligent Technologies (CONIT)*, Hubli, India, 2021, pp. 1-4, doi: 10.1109/CONIT51480.2021.9498475

[8] Kumari, Sonali & Swapnesh, Debasish & Nayak, Debasish & Swarnkar, Tripti. (2023). LOAN ELIGIBILITY PREDICTION USING MACHINE LEARNING: A COMPARATIVE APPROACH. 3. 48-54.

[9] Muhammad, I., Dahlia, R., Muhammad Ifan Rifani Ihsan, Lisnawanty, & Rabiatus Sa'adah. (2024). Performance Analysis of Ensemble Learning and Feature Selection Methods in Loan Approval Prediction at Banks. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, *3*(2), 557–564. https://doi.org/10.59934/jaiea.v3i2.426

[10] Sangeetha, D & Baba, Mohammed & Jagadish, Dr & Sriharsha, Sriram & Nikhil, Shravan. (2023). Loan Risk Prediction Using Machine Learning Algorithm.

[11] "Loan Eligibility Dataset." Kaggle,15-Aug-2020. [Online]Available:https://www.kaggle.com/datasets/vikasukani/loan-eligibledataset [Accessed 9-Feb-2024].

[12] Hussein, A.S., Li, T., Yohannese, C.W. *et al.* A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE. *Int J Comput Intell Syst* **12**, 1412–1422 (2019). https://doi.org/10.2991/ijcis.d.191114.002

[13] D. M. Powers, (2020), "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv preprint arXiv:2010.16061 . https://doi.org/10.48550/arXiv.2010.16061

[14] J.M. Hilbe. 2011, "Logistic Regression." International encyclopedia of statistical science. Vol 1: pp. 15-32.

[15] A. Saini. "Logistic Regression | What is Logistic Regression and Why do we need it?" 26-Aug-2021[Online] Available: https://www.analyticsvidhya.com/blog/2021/08/conceptual-understandingof-logistic-regression-for-datascience-beginners/#h2_5 [Accessed: 28-Jan-2022] \

[16] L.K. Ramasamy, S. Kadry, Y. Nam, & M.N. Meqdad. 2021,"Performance analysis of sentiments in Twitter dataset using SVM models. International Journal of Electrical and Computer Engineering (IJECE). Vol. 11, No. 3, pp.22752284 https://doi.org/10.11591/ijece.v11i3.

[17] R. Kunchhal. "Mathematics Behind SVM | Math Behind Support Vector Machine." 28-Dec-2020. [Online]. Available:https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behindsvm/ [Accessed: 27-Jan-2022]

[18] K. Yadav, and R. Thareja. 2019, "Comparing the performance of naive bayes and decision tree classification using R." International Journal of Intelligent Systems and Applications., Vol.11(12), p.11. DOI:10.5815/ijisa.2019.12.02

[19] Ramya, K., Teekaraman, Y. & Kumar, K.A.R. Fuzzy-Based Energy Management System With Decision Tree Algorithm for Power Security System. *Int J Comput Intell Syst* **12**, 1173–1178 (2019). https://doi.org/10.2991/ijcis.d.191016.001

[20] C. S. Reddy, A. S. Siddiq and N. Jayapandian, "Machine Learning based Loan Eligibility Prediction using Random Forest Model," *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2022, pp. 1073-1079, doi: 10.1109/ICCES54183.2022.9835875

[21] Z. Tian, J. Xiao, H. Feng, & Y. Wei. 2020, "Credit risk assessment based on gradient boosting decision tree." Procedia Computer Science. Vol.174, pp.150-160. https://doi.org/10.1016/j.procs.2020.06.070

[22] "Bagging vs Boosting in Machine Learning." GeeksforGeeks. 07,Jul2021[Online].Available:https://www.geeksforgeeks.org/bagging-vsboosting-in-machine-learning/ [Accessed 28-Jan-2022].