# Loan Approval Prediction Using KNN Approach

MAHAMODUL HASAN MAHADI
Computer Science
American International University
Bangladesh
Dhaka, Bangladesh
mahamodulhasanmahadi@gmail.com

MOHAMMAD MUSHFIQ US SALEHEEN
Computer Science
American International University
Bangladesh
Dhaka, Bangladesh
aurpon10@gmail.com

ZERIN TASNIM
Computer Science
American International University
Bangladesh
Dhaka, Bangladesh
zerin.tasnim022@gmail.com

***Abstract*— Loan approval in financial organizations is one of the challenges that affect the operational financial process due to the inaccurate estimation or the lack of information. Financial institutions need advanced, current, and customized predictive analytics to protect themselves from the frustrating fraudster. Artificial Intelligence, Machine learning and statistical methods are in high demand from data scientists and statisticians who understand them; thus, the demand for them is growing recently. The alarming rate by which loan beneficiaries default banks have course a lot of losses among many banks, and deprived many potential beneficiaries of access to the loan. This fallacy leads to inefficient and/or inaccurate management of loans in banks, and sadly, many banks have close down and have not yet realized that the labor-intensive approaches to loan management are not efficient enough. The trend has caused many banks workers to lose their job. The traditional ways of detecting fraud in bank loan management are not effective because the credit officer can easily be manipulated and not even discovered many loan defaulters. Therefore, this paper used K-Nearest Neighbors Algorithm to detect loan fraud in bank loan management to avoid loan defaulter manipulate the officer in charge of loan administration. Finally, the results of proposed method in terms of accuracy and speed have been compared and evaluated with other methods. A loan credit dataset of 600 customers in a microfinance bank was used in this study. This promising solution makes fraud detection easier, and as well provide support to the bank to detect fraud in loan management.**

***Keywords*— *Artificial Intelligence, Machine learning, Fraud Detection, Electronic Banking, KNN Algorithm***

## I. INTRODUCTION

In recent times, the population of loan applications is of increase since many depend on it for different reasons [1]. Research has shown that people do not have access to a loan in banks or from other means dues to defaulters that are not willing to pay back what they have collected. This singular act deprived potential beneficiaries the opportunity of obtaining a loan [2]. The inability of banks to manage loans by creditors efficiently can lead to credit. The continuous period of careless and unsuitable lending is called a credit crisis, thus cause losses for banks and lending societies [3]. Credit risk in the banking industry is one of the most important issues; thus, valuation of loan defaulters and risk has been given attention in the recent years [4]. Before the era of information technology, detecting fraudsters could only be done manually and the method is over simple and tedious to operate[5], thus the detection of any kind of fraudsters difficult to identify. In modern society, there are different types of fraudulent practices in financial operations, bank loan and credit administration is among them. This call for an urgent solution through modern intelligent technology. The existing bank credit administration and management for fraud detection methods are avoidance of false alarm, and the desired accuracy is not satisfactorily met. The prediction accuracy is affected by undefined fraud settings, missing data, and fraudulent duplicates. Fraud is the unlawful act of an individual to make profits at the expense of legal individuals or institutions, thus fraud is not an error by an individual. Economic sabotage can be as a result of the effect of financial sabotage. Band credit fraud, cooperative society fraud, pension fraud, bankruptcy fraud, tax evasion, counterfeit fraud, credit card fraud, money laundering, among other are examples of financial sabotage [6]. These fraudulent acts have negative impact on nation's economy, and different methods have been employed with little or no effect and with deficiencies. Therefore, Artificial Intelligence (AI) can be used to provide a more efficient and reliable solution to fraud detection in financial institutions[6]. Theoretically, statistics is a subdivision of mathematics while machine learning is a subfield of AI [7]. However, many think they only need detailed knowledge of one to be a predictive modeler. This misconception leads to mistakes and inefficient models, and sadly, many industries have not yet realized that the mathematics behind the model is just as important, if not more important, than the computer science needed to implement it. However, some industries have and this paper will move further along in this direction. Although different efforts have been put in place to combat financial fraud, which includes bank fraud detection, credit card fraud detection, credit defaulter and fraud strike force teams, fraud is still extensive. Therefore, these require new and innovative approaches to reduce financial losses in banks loan detection. A moneylender is called a loan defaulter if he/she refuses to pay the loan at an appropriate time or an approved time [3]. Also, when a moneylender refused to meet the legal condition that bond by laws of a loan agreeing by the promissory note, and not meets the obligations loan agreements. The declined in making agreed instalment paybacks by defaulters is term to be a non-performing loan depending on the type of loan taken maybe in three months or 180 days. A moneylender can be called loan defaulter if the interest or the principal of a given

loan is not paid as when due depending on the type of loan taking whether 90 days or 189 days of non-payment. The terms loans define whether a loan is a non-performing loan or not and base on the subsisting agreement conditional based on promissory notes and agreements. This study aims to apply machine learning algorithms for loan and credit detection to identify fraud and abuse in loan management and administration.

### A. K-Nearest Neighbor Algorithm

K-Nearest Neighbor is a supervised machine learning algorithm because the data passed to it is labelled. It is a non-parametric method because the classification of test data points is based on the closest training data points instead of considering the dimensions of the dataset. It is used to solve classification and regression tasks. In the classification technique, it classifies the objects based on the k closest training examples in the feature space. The working principle behind KNN assumes that the same data points are in the same environment. It reduces the effort of building a model, adjusting a set of parameters, or making more assumptions. It captures the idea of proximity based on a mathematical formula called Euclidean distance.

CHOICE OF THE PARAMETER K: An object to be classified is assigned to the respective class that represents the greater number of its nearest neighbors. If k is 1, then the data point is placed in the category containing only one nearest neighbor. Given a new input data point, the distances between those points and all data points in the training dataset are calculated. Based on the distances, the training set data points with shorter distances from the test data point are considered the nearest neighbors of our test data. Finally, the test data point is placed into one of its nearest neighbor classes. Therefore, the classification of the test data point depends on the classification of its nearest neighbors. Choosing the value of K is the crucial step in implementing the KNN algorithm. The value of K is not fixed and varies for each record depending on the type of record. When the value of K is smaller, the stability of the prediction is lower. In the same way, if we increase its value, the ambiguity will be reduced, resulting in smoother borders and increasing stability. With KNN, the assignment of a new data point to a category depends entirely on the K's value. K represents the number of closest training data points in the vicinity of a given test data point, and then the test data point is assigned to the class containing the highest number of nearest neighbors.

### B. Logistic Regression

Logistic regression is a supervised machine learning technique used in classification tasks. Logistic regression uses an equation similar to linear regression, but the logistic regression result is a categorical variable while it is a value for other regression models. Binary outcomes can be predicted from the independent variables. The result of the dependent variable is discrete. Logistic regression uses a simple equation that shows the linear relationship between the independent variables. These independent variables along with their coefficients are linearly combined to form a linear equation that is used to predict the output. The equation used by the basic logistic model. The logistic function is used here to suppress the result value between 0 and 1. The logistic function can also be called a sigmoid function or a cost function. The logistic function is a shaped curve that takes the input and changes it to a value between 0 and 1.

### C. Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

### D. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

## II. LITERATURE REVIEW

AI In several industries today, including banking, retail, intrusion detection, bioinformatics, and logical data analysis, data mining techniques are being used. Data mining techniques are very helpful in the banking industry since they improve data analysis and assist create decisions that are correct. As a result, authors in [8] have examined this topic and introduced the use of several data mining techniques, such as Random Forest, Decision Tree, Bayes classification, Bagging, and Boosting, in financial data analysis and decision making. In order to reduce manual errors for the bank, the authors [8] recommended scoring-based loan default risk analysis. Decision trees are preferred by banks for credit risk assessments because they are transparent white-box models that are simple to understand and trace. Additionally, Boosting can be used to increase the decision tree's effectiveness. Despite the fact that the authors of [8] provided a clear justification for the techniques they utilized, the work is devoid of experimental data, in-depth analysis, or model evaluation. The authors of [9] provided a thorough evaluation framework for the loan classification models used by banks in order to assess the currently in use classification models. The suggested evaluation technique is based on multiple criteria decision making (MCDM), in which the high-rank prediction model is chosen by comparing the performance of the various models to a set of performance indicators.

Predictive analytics heavily relies on the process of data mining. Data mining is the process of obtaining information or specifics from a vast amount of data. Despite the fact that data mining is a part of knowledge discovery in databases (KDD), data mining is most frequently connected with KDD. The goal of data mining functions is to recognize the various patterns that can be found in data mining jobs. Models are derived from datasets using data mining techniques, where a dataset is a collection of details. Data mining techniques learn from datasets, that is, they learn to predict the important result

from a specific input [10]. This type of knowledge acquisition has no impact on the workstations' ability to hold onto data, but it does change how they function so that future improvements can be made [11].

The decision tree technique was used by the authors of [12] to conduct research on the retail bank's credit risk. The study was carried out to help the banking industry analyze loan data for credit decision-making. By providing a wealth of information to the credit decision-making process, this research was carried out to support the banking industry. This is accomplished by reducing the amount of money and time wasted on loan review, as well as the degree of vulnerability experienced by loan authorities, by providing them with knowledge gleaned from prior loan data via a decision tree.

The authors of [13] proposed creating a model to forecast potential business sectors in retail banking. For the research, records from Bangladesh's rural and urban areas as well as records of business clients of a retail bank were used. The primary transactional determinants of clients were broken down using these records, and a prototype for likely subdivisions in the retail bank was predicted. Weka was used to execute the decision tree data mining technique after it had been used to analyze the problems required to develop the model. The creation of a Credit Scoring Model for use by Sudanese Banks was the paper's main objective. The Decision Tree (DT) and Artificial Neural Network (ANN) classifications of data mining were the two that were selected (ANN). Then, as feature selection methods, Generic Algorithm (GA) and Principal Component Analysis (PCA) were employed. German and Sudanese credit datasets were used for the approaches' evaluation. The results of the categorization showed that ANN typically outperforms DT. Additionally, it was discovered that GA is superior to PCA approach for feature selection. The accuracy of the German data set was 80.67%, compared to the Sudanese data set's accuracy of 69.74%. Last but not least it was observed that ANN outperformed DT as well as PCA-DT and GA-DT, two of its hybrid models.

Similarly, the authors of [14] suggested data mining as a creative method for classifying credit risk in the banking industry. To predict the state of loans, the data for this model came from the banking industry. The predicted models were developed using three algorithms: J48, BayesNet, and Naive Bayesian. Weka was the application that was utilized for implementation and testing. The study's findings indicated that j48 was the most accurate. This study examined how five classifiers for credit risk prediction accuracy behaved under different types of interference and how classifier ensembles could increase precision. On the basis of four credit datasets, the findings are then contrasted with each classifier's performance in terms of prediction accuracy at various attribute noise levels. The outcomes of the experiment indicate that it is feasible to increase prediction accuracy by employing an ensemble of classifiers.

In 2016 some researchers explored ways to predict how a bank will grant a loan. They provided a model that makes use of machine learning tools like neural networks and SVM. This evaluation of the literature helped us conduct our research and create a trustworthy bank loan prediction model [15].

A study to forecast whether or not a bank will grant a loan to a customer was offered by the authors [16]. The model's objective was classification hence it was developed using logistic regression and the sigmoid function. Two data sets one for training and the other for testing made up the dataset used for research and prediction, which was collected from Kaggle. The data must first be cleaned up in order to prevent missing values in the data set. The models were then evaluated using performance metrics including sensitivity and specificity. The final result showed an accuracy of 81% for the model. The model was slightly superior because it took into account factors other than checking account information, which indicates a customer's wealth and should be taken into account when accurately calculating the probability of loan default, such as a customer's age, purpose, credit history, credit amount, and credit duration. As a result, the appropriate clients to target for loan giving may be easily identified using a logistic regression approach by assessing the likelihood of default on a loan.

The main objective of this paper by the authors of [17] was to reduce the risk element involved with selecting a trustworthy individual to assign the loan in order to save time and money for the bank. This paper was divided into four parts. Data collection, model comparison using the gathered data, system training using the most promising model, and testing are the first three steps. In this article, they used machine learning techniques like classification, logistic regression, decision trees, and gradient boosting to forecast loan data. The decision tree algorithm was found to be the most accurate when compared to other algorithms, with an accuracy of 82 percent. It was effective since the classification problem results were improved. It was highly user-friendly, easy to install, and produced results that could be understood.

## III. Proposed Loan approval prediction model

A credit dataset that contains 600 of a microfinance bank was used with feature extraction that target attribute is the default status. Feature extraction target defaulter status was used to select twelve (12) attributes that are related to loan disbursement. For effective pre-processing before the successive operation, the twelve-feature selection was determined. The simulation was performed using Python programming language for fraud detection. The proposed model developed was used for training and testing of the bank loan dataset to show the effectiveness of the system. Accuracy, true positive rate, and false-positive rate were used for performance evaluation in this study.

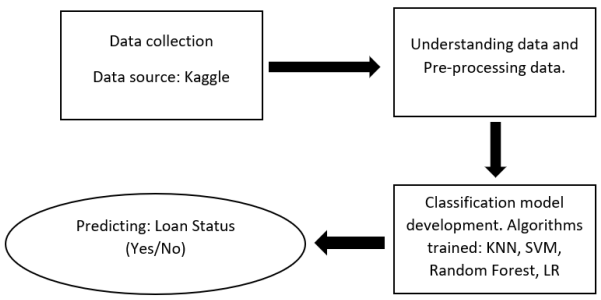


Fig 01: Working Model.

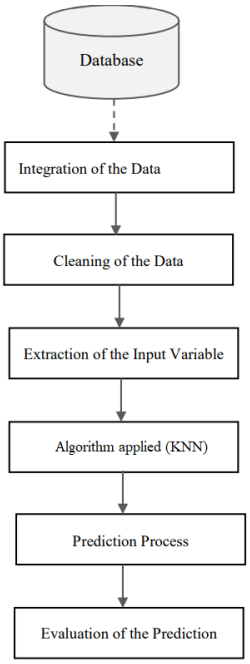


Fig 02: The Process to Predict Bank Loan Fraud Detection using KNN

## IV. Dataset Description

Dataset selected from Kaggle. This dataset contains 13 features.

Loan: A unique id.
Gender: Gender of the applicant Male/female.
Married: Marital Status of the applicant, values will be Yes/ No.
Dependents: It tells whether the applicant has any dependents or not.
Education: It will tell us whether the applicant is Graduated or not.
Self_Employed: This defines that the applicant is self-employed i.e. Yes/ No.
ApplicantIncome: Applicant income.
CoapplicantIncome: Co-applicant income.
LoanAmount: Loan amount (in thousands).
Loan_Amount_Term: Terms of loan (in months).
Credit_History: Credit history of individual's repayment of their debts.
Property_Area: Area of property i.e. Rural/Urban/Semi-urban.
Loan_Status: Status of Loan Approved or not i.e. Y- Yes, N-No.

| | Loan_ID | Gender | Married | Dependen | Education | Self_Empl | Applicantl | Coapplica | LoanAmo | Loan_Am | Credit_His | Property_ | Loan_Status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | LP001002 | Male | No | 0 | Graduate | No | 5849 | 0 | | 360 | 1 | Urban | Y |
| 3 | LP001003 | Male | Yes | 1 | Graduate | No | 4583 | 1508 | 128 | 360 | 1 | Rural | N |
| 4 | LP001005 | Male | Yes | 0 | Graduate | Yes | 3000 | 0 | 66 | 360 | 1 | Urban | Y |
| 5 | LP001006 | Male | Yes | 0 | Not Gradu | No | 2583 | 2358 | 120 | 360 | 1 | Urban | Y |
| 6 | LP001008 | Male | No | 0 | Graduate | No | 6000 | 0 | 141 | 360 | 1 | Urban | Y |
| 7 | LP001011 | Male | Yes | 2 | Graduate | Yes | 5417 | 4196 | 267 | 360 | 1 | Urban | Y |
| 8 | LP001013 | Male | Yes | 0 | Not Gradu | No | 2333 | 1516 | 95 | 360 | 1 | Urban | Y |
| 9 | LP001014 | Male | Yes | 3 | Graduate | No | 3036 | 2504 | 158 | 360 | 0 | Semiurbar | N |
| 10 | LP001018 | Male | Yes | 2 | Graduate | No | 4006 | 1526 | 168 | 360 | 1 | Urban | Y |
| 11 | LP001020 | Male | Yes | 1 | Graduate | No | 12841 | 10968 | 349 | 360 | 1 | Semiurbar | N |
| 12 | LP001024 | Male | Yes | 2 | Graduate | No | 3200 | 700 | 70 | 360 | 1 | Urban | Y |
| 13 | LP001027 | Male | Yes | 2 | Graduate | Yes | 2500 | 1840 | 109 | 360 | 1 | Urban | Y |
| 14 | LP001028 | Male | Yes | 2 | Graduate | No | 3073 | 8106 | 200 | 360 | 1 | Urban | Y |
| 15 | LP001029 | Male | No | 0 | Graduate | No | 1853 | 2840 | 114 | 360 | 1 | Rural | N |
| 16 | LP001030 | Male | Yes | 2 | Graduate | No | 1299 | 1086 | 17 | 120 | 1 | Urban | Y |
| 17 | LP001032 | Male | No | 0 | Graduate | No | 4950 | 0 | 125 | 360 | 1 | Urban | Y |
| 18 | LP001034 | Male | No | 1 | Not Gradu | No | 3596 | 0 | 100 | 240 | | Urban | Y |
| 19 | LP001036 | Female | No | 0 | Graduate | No | 3510 | 0 | 76 | 360 | 0 | Urban | N |
| 20 | LP001038 | Male | Yes | 0 | Not Gradu | No | 4887 | 0 | 133 | 360 | 1 | Rural | N |
| 21 | LP001041 | Male | Yes | 0 | Graduate | Yes | 2600 | 3500 | 115 | | 1 | Urban | Y |

Fig 03: Dataset

## V. Feature Extraction and Selection

Feature extraction is a part of the dimensionality reduction process, in which, an initial set of the raw data is divided and reduced to more manageable groups. So, it will be easier to process. The most important characteristic of these large data sets is that it has a large number of variables. For this project dataset there is 13 variables. These variables require a lot of computing resources to process. So, Feature extraction helps to get the best feature from those big data sets by selecting and combining variables into features, thus, effectively reducing the amount of data. These features are easy to process, but still able to describe the actual data set with accuracy and originality.

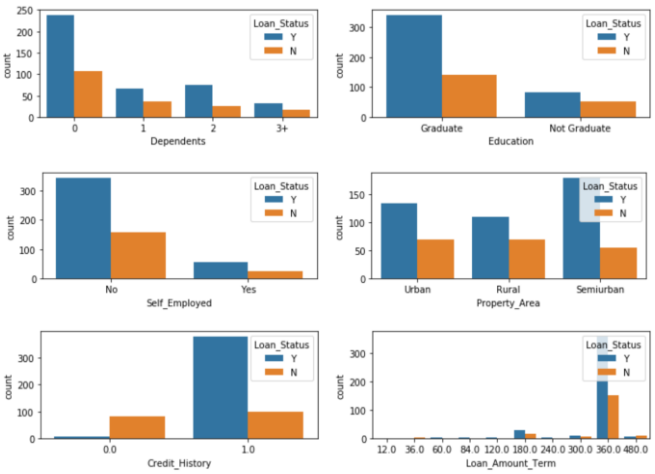


Fig 04: Outlier detection using bar plots techniques.

Loan Approval Status: About 2/3rd of applicants have been granted loan.
Sex: There are more Men than Women (approx. 3x).
Marital Status: 2/3rd of the population in the dataset is Marred; Married applicants are more likely to be granted loans.
Dependents: Majority of the population have zero dependents and are also likely to accepted for loan.
Education: About 5/6th of the population is Graduate and graduates have higher proportion of loan approval.
Employment: 5/6th of population is not self-employed.
Property Area: More applicants from Semi-urban and also likely to be granted loans.
Applicant with credit history is far more likely to be accepted.
Loan Amount Term: Majority of the loans taken are for 360 Months (30 years).

## VI. Data Understanding and Data Selection

The bank loan prediction system dataset comes from the Kaggle competition and includes applicants of various ages and genders. The data set contains twenty attributes, such as education, marital status, income, assets, and so on. There are total of 600 applicant records with the values of their relevant attributes in categorical and numerical data. We handle the missing value and normalize the data through pre-processing and feature engineering so that we may use it in an ML algorithm. The dataset is separated into two sections: training and testing. The model is trained using machine learning methods and forecasts the system using test data, as detailed in the following section. In this dataset,

Data columns (total 13 columns):

| | |
|---|---|
| Loan_ID | 614 non-null object |
| Gender | 601 non-null object |
| Married | 611 non-null object |
| Dependents | 599 non-null object |
| Education | 614 non-null object |
| Self_Employed | 582 non-null object |
| ApplicantIncome | 614 non-null int64 |
| CoapplicantIncome | 614 non-null float64 |
| LoanAmount | 592 non-null float64 |
| Loan_Amount_Term | 600 non-null float64 |
| Credit_History | 564 non-null float64 |
| Property_Area | 614 non-null object |
| Loan_Status | 614 non-null object |

dtypes: float64(4), int64(1), object(8)

As Loan_ID is completely unique and not correlated with any of the other column so this will be dropped out from the dataset. To test the correlation between data attributes to find the most noteworthy feature in the prediction process. For this purpose, we use a heat map to visualize the correlation of the variables. Figure 5 depicts the heat map for the collected data attributes. From the heat map in figure 5, we can easily notice the most important feature for loan prediction. Notably, Loan_ID has been removed from the heat map, as it has no impact on the prediction process.

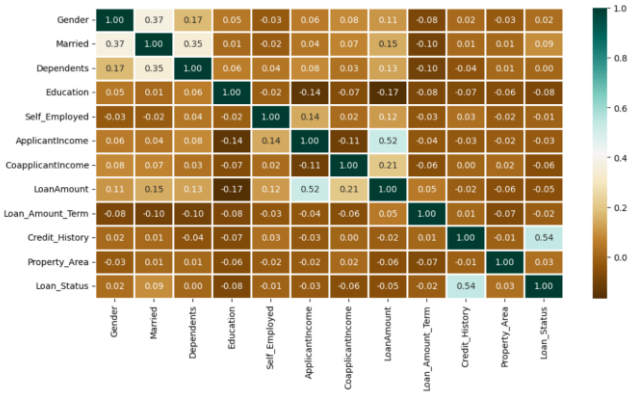


Fig 05: Representing the correlation between attributes using the heat map.

## VII. Training and Classification

After performing feature extraction and data selection, we have trained and tested four algorithms on the data. Classifications of the data sets are done on the basis of specific properties possess by the sample variable that the capable to classify them, and each sample variable is assigned a malignant or benign class. Classification is principally done by making predictions based on known sample data that has been learned from training data. Designed algorithm is first trained on the known data labels and further uses this learning to predict the class labels for the new unknown set of data sample. We train the classifier with known sample data in a training dataset and check its performance by examining the test dataset, which consists of the unknown sample used to predict its class label K Neighbors Classifier is a supervised, instance-based learning classifier which learns from the labelled data samples. K folds cross validation technique is used for training data. In this technique, the original sample is divided into k equivalent size subsamples and one subsample is used for validating the model, while the k-1 remaining subsamples are utilized as training data. After that this cross-validation process is recurring k times, with every of the k subsamples just used one time as the validation data. It works in loop manner. In this study we set the value of k = 1, 2, 3, 5, 8, 10, 15, 20. The second algorithm is Logistic regression. Logistic regression is a statistical technique that is used to predict the probability of binary response variables. It is used when our label(y) is a binary response variable in 1 or 0, yes or no, etc. It is a basic and popular algorithm to solve a classification problem. The third algorithm is a Decision Tree classifier. A decision tree is a popular algorithm that is used to build classification models. The models are built in the form of a tree-like structure. Each node in the tree indicates a test on a variable, and each branch descending from that node indicates one of the possible values for that attribute. The fourth algorithm is Random Forest Classifier. Random Forest is a classification algorithm that consists of many decision trees. It tries to create an uncorrelated forest of trees by using feature randomness and Bagging. The prediction of an uncorrelated forest of trees is more accurate than that of any individual tree.

## VIII. Result and Data Analysis

**For K value = 1,**

Training result

```
Accuracy score of  RandomForestClassifier = 98.04469273743017
Accuracy score of  KNeighborsClassifier = 100.0
Accuracy score of  SVC = 68.71508379888269
Accuracy score of  LogisticRegression = 79.88826815642457
```

Testing result

```
Accuracy score of  RandomForestClassifier = 97.5
Accuracy score of  KNeighborsClassifier = 100.0
Accuracy score of  SVC = 70.41666666666667
Accuracy score of  LogisticRegression = 82.91666666666667
```

**For K value = 2**

Training result

```
Accuracy score of  RandomForestClassifier = 98.04469273743017
Accuracy score of  KNeighborsClassifier = 79.05027932960894
Accuracy score of  SVC = 68.71508379888269
Accuracy score of  LogisticRegression = 79.88826815642457
```

Testing result

```
Accuracy score of  RandomForestClassifier = 97.5
Accuracy score of  KNeighborsClassifier = 79.58333333333333
Accuracy score of  SVC = 70.41666666666667
Accuracy score of  LogisticRegression = 82.91666666666667
```

**For K value = 3**

Training result

```
Accuracy score of  RandomForestClassifier = 98.04469273743017
Accuracy score of  KNeighborsClassifier = 78.49162011173185
Accuracy score of  SVC = 68.71508379888269
Accuracy score of  LogisticRegression = 79.88826815642457
```

Testing result

```
Accuracy score of  RandomForestClassifier = 97.5
Accuracy score of  KNeighborsClassifier = 76.66666666666667
Accuracy score of  SVC = 70.41666666666667
Accuracy score of  LogisticRegression = 82.91666666666667
```

**For K value = 5**

Training result

```
Accuracy score of  RandomForestClassifier = 98.04469273743017
Accuracy score of  KNeighborsClassifier = 71.22905027932961
Accuracy score of  SVC = 68.71508379888269
Accuracy score of  LogisticRegression = 79.88826815642457
```

Testing result

```
Accuracy score of  RandomForestClassifier = 97.5
Accuracy score of  KNeighborsClassifier = 72.5
Accuracy score of  SVC = 70.41666666666667
Accuracy score of  LogisticRegression = 82.91666666666667
```

**For K value = 8**

Training result

```
Accuracy score of  RandomForestClassifier = 98.04469273743017
Accuracy score of  KNeighborsClassifier = 70.6703910614525
Accuracy score of  SVC = 68.71508379888269
Accuracy score of  LogisticRegression = 79.88826815642457
```

Testing result

```
Accuracy score of  RandomForestClassifier = 97.5
Accuracy score of  KNeighborsClassifier = 70.41666666666667
Accuracy score of  SVC = 70.41666666666667
Accuracy score of  LogisticRegression = 82.91666666666667
```

**For K value = 10**

Training result

```
Accuracy score of  RandomForestClassifier = 98.04469273743017
Accuracy score of  KNeighborsClassifier = 70.94972067039106
Accuracy score of  SVC = 68.71508379888269
Accuracy score of  LogisticRegression = 79.88826815642457
```

Testing result

```
Accuracy score of  RandomForestClassifier = 97.5
Accuracy score of  KNeighborsClassifier = 70.83333333333334
Accuracy score of  SVC = 70.41666666666667
Accuracy score of  LogisticRegression = 82.91666666666667
```

**For K value = 15**

Training result

```
Accuracy score of  RandomForestClassifier = 98.04469273743017
Accuracy score of  KNeighborsClassifier = 70.39106145251397
Accuracy score of  SVC = 68.71508379888269
Accuracy score of  LogisticRegression = 79.88826815642457
```

Testing result

```
Accuracy score of  RandomForestClassifier = 97.5
Accuracy score of  KNeighborsClassifier = 70.0
Accuracy score of  SVC = 70.41666666666667
Accuracy score of  LogisticRegression = 82.91666666666667
```

**For K value = 20**

Training result

```
Accuracy score of  RandomForestClassifier = 98.04469273743017
Accuracy score of  KNeighborsClassifier = 68.71508379888269
Accuracy score of  SVC = 68.71508379888269
Accuracy score of  LogisticRegression = 79.88826815642457
```

Testing result

```
Accuracy score of  RandomForestClassifier = 97.5
Accuracy score of  KNeighborsClassifier = 69.58333333333333
Accuracy score of  SVC = 70.41666666666667
Accuracy score of  LogisticRegression = 82.91666666666667
```

We have randomly chosen the data construct the training set. We have plotted a graph to check the error rate against the k-value and random classifier and we have been discussing different values of K from K = 1 to 20 while changing the training and testing size. When the k value is 1, 2 and 3 the KNN test and Train set result was so high. From 1 it was 100.0, which is the highest. But gradually it is dropping but for the random classifier the value is stable in both set result. From k value 5 to 20 the KN classifier's value is dropping all the way around 68.7. But for the Random forecast classifier the value was stable all along. So, the less K value the better result in K-neighbor Classifier value's and this results fluctuates between 100.0% and 68.71%. The best performance is obtained when K value is lower. Also, SVM and Logistic Regression are also performing good for this dataset.

## IX. Discussion

The fundamental phases in the analytical process include data preparation, data processing, missing value imputing, exploratory data analysis, model creation, and model assessment. By looking at the outcomes, the following observations concerning loan approval are made. 100.0 is the greatest attainable accuracy by using the KNN model with the given dataset. Due to the applicant's very low responsibility and little likelihood of repaying the loan, those with a "0" credit history were not successful in getting their loan applications accepted. Additionally, persons who had fewer dependents, a higher salary, and a smaller loan request were approved. The least significant characteristics were thought to be the applicant's gender and educational background. There are 12 features in the dataset, and we found that Credit_History is the most important feature for the prediction of loans. The pre-processing starts with data understanding, data cleaning, outlier detection, and removal. Four machine-learning algorithms were trained and tested in the proposed prediction model on the data: KNN, logistic regression, SVM, and Random Forest. Random Forest showed better performance than the others with around 95%-99% accuracy, while SVM and logistic regression got 69%-70% and 79%-82% accuracy, respectively and KNN got 69%-79% accuracy. The recorded results have been validated using the ROC curve. In addition, the proposed model was compared with the other model. The possible future directions of this work will be to acquire a realistic dataset with more prediction features to improve the prediction. Besides, the prediction accuracy must be improved. Such improvement can be achieved through applying feature extraction applying hybrid machine learning algorithms.

## X. Conclusion

The loss of non-creditworthy customers has created a huge amount of loss for banks and other sectors, thus fraud detection has become useful in the financial segments. But the detection and prediction of fraud in financial sectors are very difficult due to the diversity of applicant behaviors. This study provided an intelligent model based on KNN for detecting loan fraud in a highly competitive market for credit leaden limits management. KNN simplifies how banks would detect loan fraud within credit management and will make an efficient judgment in the event of a reduction in loaning supply if faced with a negative liquidity shock. Hence,

concentrate on the primary goal of increasing bank profits. The results show that KNN greatly detect fraud among loan lenders and loan administrators. Therefore, the bank profit is increased by implementing the advised loan choice based on real facts. The results reveal that our proposed method outperforms other state-of-the-art methods using real transaction data from a financial institution. Future work could apply a genetic algorithm for better feature selection, which would improve the system's performance, as well as a hybrid technique for a better result.

## ACKNOWLEDGMENT

The authors would like to acknowledge to Ajay M for supporting us with the Loan Approval Prediction dataset that we used for this project collected from Kaggle.

Dataset link:

https://www.kaggle.com/code/ajaymanwani/loan-approval-prediction/notebook [17].

## REFERENCES

[1] Gupta, A., Pant, V., Kumar, S., & Bansal, P. K. (2020, December). Bank Loan Prediction System using Machine Learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE.

[2] Kemalbay, G., & Korkmazoğlu, Ö. B. (2014). Categorical principal component logistic regression: a case study for housing loan approval. Procedia-Social and Behavioral Sciences, 109, 730-736.

[3] Sánchez, J. S., Mollineda, R. A., & Sotoca, J. M. (2007). An analysis of how training data complexity affects the nearest neighbor classifiers. Pattern Analysis and Applications, 10(3), 189-201.

[4] Raniszewski, M. (2010, September). Sequential reduction algorithm for nearest neighbor rule. In International Conference on Computer Vision and Graphics (pp. 219-226). Springer, Berlin, Heidelberg.

[5] Coomans, D., & Massart, D. L. (1982). Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules. Analytica Chimica Acta, 136, 15-27.

[6] Young, M. (1989, January 01). The Technical Writer's Handbook (1989 edition). Retrieved December 23, 2022, from https://openlibrary.org/books/OL24815867M/The_technical_writer's_handbook

[7] Harrison, O. (2019, July 14). Machine learning basics with the K-nearest neighbors algorithm. Retrieved December 23, 2022, from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 ZHOU, P.Y., CHAN, K.C., OU, C.X., and CHAWAN, P.M. Corporate communication network and stock price movements: insights from data mining. IEEE Transactions on Computational Social Systems, 2018, 5(2): 391-402.

[8] YONGMING, S., and PENG, Y. A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction. IEEE Access, 2019, 7: 84897-84906.

[9] Stevens, D. (2014). Predicting real estate price using text mining automated real estate description analysis. International Journal of Advanced Research in computer and communication engineering, 7(4), 3-6.

[10] Witten, H.; and Frank, E. (2017). Data mining: practical machine learning tools and techniques. San Francisco, CA: Morgan Kaufmann.

[11] Sudhakar, M; and Reddy, K. (2016). Two step credit risk assessment model for retail bank loan applications using decision tree data mining technique. International Journal of Advanced Research in Computer Engineering and Technology, 5(3), 705-718

[12] Rafiqul, I; and Ahsan H. (2015). A data mining approach to predict prospective business sector for lending in retail banking using decision tree. International Journal of Data Mining & Knowledge Management Process, 5(2), 13-18

[13] Jafar, A. (2016). Mohammed T. Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications: An International journal, 3(1), 1-8.

[14] Kumar, Arun, Garg Ishan, and Kaur Sanmeet. "Loan approval prediction based on machine learning approach." IOSR J. Comput. Eng 18.3, 18-21, 2016.

[15] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learni ng Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 490-494, 2020.

[16] Supriya, P., Pavani, M., Saisushma, N., Kumari, N. V., & Vikas, K. (2019). Loan prediction by using machine learning models. International Journal of Engineering and Techniques, 5(22), 144-148.

[17] Ajaymanwani. (2019, January 17). Loan approval prediction. Retrieved December 23, 2022, from https://www.kaggle.com/code/ajaymanwani/loan-approval-prediction/notebook