**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**

**IMPACT FACTOR: 7.056**

***IJCSMC, Vol. 10, Issue. 12, December 2021, pg.71 – 82***

# A New Prediction Approach for Preventing Default Customers from Applying Personal Loans Using Machine Learning

**Mohamed H. Khedr[1]; Nesrine A. Azim[2]; Ammar M. Ammar[3]**

[1]Department of Software Engineering, Faculty of Graduate Studies for Statistical Research (FGSSR), Cairo University, Egypt

[2]Department of Information Systems and Technology, FGSSR, Cairo University, Egypt

[3]Department of Computer Science, FGSSR, Cairo University, Egypt

[1] muhammedkhedr1982@gmail.com; [2] nesrinealiazim79@hotmail.com; [3] ammar@cu.edu.eg

**DOI: 10.47760/ijcsmc.2021.v10i12.009**

***Abstract— In the Egyptian banking industry, loan officers use pure judgment to make personal loan approval decisions. In this paper, we develop a new predictive method for default customers' loans using machine learning. The new predictive method uses the available personal data and historical credit data to evaluate the credit trust-worthiness of customers to obtain loans. We used the ABE dataset for training and testing, as we used 10 features from the application form and i- score report class that could give great help to credit officers for taking the right decision through avoiding customer selection using random techniques. The collected dataset was analysed by using various machine learning classifiers based on important selected features, to obtain high accuracy. We compared the performance of several machine learning classifiers before and after feature selection. We have found that in terms of high accuracy, the most important features are (activity – income – loan) and in terms of better performance the decision tree classifier has surpassed any other machine learning classifier with significant prediction accuracy of almost 94.85%.***

***Keywords— Machine learning, Personal loans, Credit risk, default customers, predictive method***

## I. Introduction

The history of banking began in Mesopotamia more than 3,000 years ago, where Sumer began to practice interest-bearing loans. Loans were provided to farmers and merchants who transported goods between cities. Also, there is evidence that there was lending activity in the Roman Empire, Ancient China, Ancient Egypt and India, and Ancient Greece [2],[7].

The current banking structure began in the middle ages, in the 13th and 14th centuries. The first bank was named the bank was the "Bank of Barcelona" in 1401 [9].

During the 20th century, due to the advancement of communications and computing services, the banking business has changed, and its scale and geographic distribution have also increased. Credit is defined as a time allowed to one party to provide a debt incurred for goods or services to another party. While credit risk is defined as a risk that the borrower may not be able to repay on time and it is sometimes called default risk, performance risk, or counterparty risk [4],[8].

Personal loans are specific for household or other personal purposes (credit cards loans - for the purchase of electrical or electronic equipment - car loans) [4] but not for commercial purposes (investment purposes- businesses).

Avoiding default customers of investment and enterprise loans is much easier because feasibility studies and financial statements are analysed as well as customer and company reputations in the market. For personal loans, banks must gather accurate information about their potential borrowers and monitor the accepted borrowers' performance during the loan time [8].

In Egypt, the development of a scoring system: Egyptian credit bureau (i. Score Company) provides information about the historical credit of the customers and also helps to avoid default customers. However, it is not an easy task due to the limitation of availability, and interpretation techniques of customers' data and credit history.

In this paper, we propose a new efficient approach of applying machine learning techniques to predict default customers of personal loans. The approach depends on analysing customer data in the application form and i-score report class which can help credit staff to take the right decision and decrease the credit risk. We use a dataset from 27 branches of ABE that deal with personal loans. The paper compares several methods before and after applying feature selection. The rest of this paper is organized as follows: Section 2 presents previous studies that compared the performance of using machine learning techniques to predict default customers of personal loans. Section 3 introduces a background of machine learning techniques. Section 4 presents data collection and the pre-processing that had been done on this dataset. Section 5 provides the results before and after features selection. And finally, Section 6 concludes the paper.

## II. RELATED WORK

### *A. Xiao et al (2006) [20]*

They made a comparison of the performance of Logistic regression, KNN, decision tree, SVM classifiers, MARS, and ANN, for classifying the loan applications.

A personal loan dataset was provided by German banks ( it consists of 1000 applicants), Australian banks (it consists of 700 applicants) and The third credit data is from major financial institutions in the US, where there are (it consists of 1225 applications).

The features used are including (credit history, account balances, loan purpose, loan amount, employment status, age, housing, and job).

Research findings indicate that in terms of the classification ratio the best performance classifier was Logistic regression as it gains (74.2%, 88.2%, and 63.5%).

However, it has to be noted that LDA and decision tree are significantly more accurate than any other model in predicting bad customers for German and American credit data sets.

### *B. Abdou & Pointon (2009) [1]*

They made a comparison between the performance of the Discriminant analysis model (DA), the Logistic regression model, and neural network models for classifying the loan applications.

A personal loan dataset was collected from one of the Egyptian commercial banks.

The dataset consists of 1262 personal loans with 851 good loans and 411 bad loans. Each bank customer in this dataset is linked to19 features, as loan amount, loan duration, age, gender, dependents, profession, education, house status, telephone, monthly income,CBE report, guarantees, field visit, and feasibility study, and credit card status, loans from other banks, car ownership, and marital status.

Research findings indicate that Neural Networks models had a better average correct classification rate (87.64%) which is higher than the other techniques.

### *C. Ince & Aktan (2009) [11]*

They compared the performance of (discriminant analysis- logistic regression- decision trees c5/ cart - neural networks) to classify the loan applications. A credit card dataset collected from a Turkish bank is used.

 

Each bank customer record in the data set contains nine features named: gender, age, marital status, educational level, occupation, job position, income, customer type and credit cards from the other banks. The dependent variable is the credit status (good or bad credit). The dataset is composed of 1260 customers' records.

Research findings indicate that the results show that the cart has 65.58% which is the best average correct classification rate in comparison with discriminant analysis, logistic, regression, and neural networks.

On the other hand, the neural network credit scoring model has lower type II errors (a customer with bad credit is misclassified as a customer with good credit) associated with high costs and thus has better overall credit scoring capabilities.

### *D. Yeh & Lien (2009) [21]*

They made a comparison between the performances of (discriminant analysis - logistic regression- Bayes classifier- nearest neighbour- artificial neural networks- classification trees) to classify the loan applications using a dataset from one of Taiwan's major banks and the targets are the bank's credit cardholders (25,000) records . The customer features including (Gender - Amount of the given credit- Education- Marital status- Age- History of past payment).

Research findings indicate that the artificial neural network has higher classification accuracy among all the methods with a rate (54%).

### *E. Sudhakar & Reddy (2014) [18]*

They made a comparison of the performance of (logistic regression - radial basis network- multilayer perceptron model – SVM- decision tree) for classifying the loan applications. Data used is 500 customers' data and was collected from different banks in India such as SBI, Andhra, ICICI, Homeownership, and Syndicate banks. It consists of 10 different independent features (Age of customer- Sex- Marital status- Service period- Current account- Saving account- Payment history- Occupation- Homeownership- Address time) and one dependent variable(Credit (Approved or Not).

Research findings indicate that accuracy for SVM (85.97%), decision tree (84.77%), and radial basis network (86.53%) are the best methodologies for classifying loan applications. By analysing the performance of these models on the standard dataset it is found that in the case of missing data multilayer perceptron model and logistic regression are also good. The recommendation is to develop a new integrated model that takes advantage of all five models.

### *F. Sousa & Figueiredo (2014) [16]*

They made a comparison of the performance of (Decision Tree- Neural Network) for classifying the loan applications. The data were collected from 211 individual customers from the Credit Union of the SICOOB in Brazil. The study used 26 features, such as (gender-age – level of education- birth place-place of residence-work or activity-Marital status-Years of experience inactivity/work).

Research findings indicate that the decision tree's performance for the problem is (94.08%) which is higher than the Neural Network (91.59%).

### *G. Jafar & Ahmed (2016) [12]*

They made a comparison of the performance of (decision tree (j48) - Bayes Net - naïve Bayes) for classifying the loan applications. the dataset was collected from the banking sector in Sudan (1000 instances) using 7 features (Credit history- Purpose- Gender- Credit amount- Age,-Housing- Job ) and one dependent variable (the loan Class (good - bad).

Research findings indicate that applying classification algorithms which are decision tree (j48), Bayes Net, and naïve Bayes, found that the best classifier for loan classification is decision tree (j48) algorithm (78.38%) because it has high accuracy and low mean absolute error.

**Based on the above-mentioned studies, it is useful to highlight the following points:-**

1- They did not use large datasets (not exceeding 25,000 customers).

2- Most of them did not compare the performance of a wide variety of classification algorithms.

3- They did not compare the performance of classifiers before and after feature selection.

 

## III.BACKGROUND

The Term of machine learning represents learning from experience (in this case is the previous data) to improve future performance [5].

According to learning style, machine learning algorithms can be divided into the following subcategories:

A. *Supervised learning*: the Input data has a pre-defined label such as True/False, Positive/Negative, and Spam/Not Spam, etc. a classifier is created and trained to expect the test data labels [5].
B. *Unsupervised learning*: the Input data is not classified. The classifier creation is through inferring the existing pattern or clusters in the training datasets [5].
C. *Semi-supervised learning*: The training dataset contains both labelled and unlabelled data. The classifiers are trained to learn the patterns for classifying and labelling data and making predictions [5].
D. *Reinforcement learning*: training the algorithm to represent the action in a situation to maximize the feedback signal .by using trial and error the designed classifier is not selecting the action directly but finding the most useful actions [5].
E. *Transduction*: it attempts to expect the outputs based on the training data, the training labels, and the test data. However, it has similarities to supervised learning, but it does not develop an explicit classifier [5].
F. *Learning to learn*: The classifier is trained to learn from the bias is generated in the previous stages [5].

There are different techniques of machine learning used for classification, such as K- nearest neighbour classifiers, Logistic regression, Discriminant analysis, Naive Bayesian classifier, SVM, Artificial neural networks, and decision trees.

1. *K-nearest neighbour (KNN) classifiers*: the training data (well-labelled) is fed into the classifier. When the test data is entered into the classifier, it compares both of the data. k is the most correlated data that is generated from the training set [6].
2. *Logistic regression*: it can predict discrete outcomes from a group of variables that can be continuous, discrete, dichotomous, or a mix of any of these [13].
3. *Discriminate analysis*: it is an alternative to logistic regression and it assumes that the explanatory variables for each given class of response variable, are distributed a multivariate normal distribution with a common variance-covariance matrix [13].
4. *The naive Bayes classifier*: it is a probabilistic classifier that assumes the existence or nonexistence of a specific feature of a class is unrelated to the existence or non-existence of any other feature [13].
5. *Support vector machine*: it determines an optimal separating hyper plane that separates the two classes of data points justly and is equidistant from both of them [5].
6. *Decision tree*: it is a type of tree that groups attribute by sorting them based on their values and each tree has nodes and branches. Each branch represents a value that the node can take while each node represents attributes in a group that is to be classified [6].
7. *Random Forest*: is a generalization of recursive partitioning that combines a collection of trees called an ensemble [3].
8. *Artificial neural networks*: it simulates the function of the human brain. it is a powerful tool for unknown data relationship modelling. It can recognize the complex pattern between input and output variables then predict the outcomes of the new independent input data [13].

## IV.PROPOSED WORK

A. *Dataset Description*:

We collected a dataset that provides the ABE bank customers‟ information including 122,572 records and 11 fields of customers who have applied for a personal loan to the ABE bank for five years from 2013 to 2018, with loans ranging from 12 to 60 months.

The dataset was collected from 27 branches of ABE bank which deals with personal loans including (Cairo – Qaliobia – Alexandria – Matrouh - Kafr Elsheikh- Damietta - Ismailia- port said - Behira – Giza – Dakahlia – Suez –Sharkia –Gharbia – Menofia - North of Sinai –South of Sinai – Luxor –Red sea - Fayoum – Minia –Beni swif- Sohag – Qena– Assiut – New valley – Aswan).

This dataset contained both categorical attributes and numeric ones as described below:

1- Address: the town or city in which the customer lives (categorical).
2- Activity: type of job or activity of the customer (categorical).
3- Gender: sex type of the customer (categorical)
4- Marital: Marital Status of the customer (categorical)
5- Age: age of the customer (numeric)
6- Income: the amount of customer annual income (categorical)
7- Loan: the amount of loan requested for (numeric).
8- Duration: the time of loan requested for (categorical)

9- Type: the purpose of loan requested for (categorical)
10- I-score: the customer credit score report status (categorical)
11- Status: the current credit status of a consumer (categorical)

B. *Dataset pre-processing:*
We made sure that there are no missing values exist in the dataset as it could have effects on the training process to pre-process the dataset, we used the Python library scikit-learn (Sklearn).
We have done the following steps to pre-process the collected dataset:-

1. *Feature Encoding:*
   The dataset contains values in the form of strings and numbers so; we converted every categorical value into numerical ones as the machine learning algorithms cannot run datasets with string values.
   We used Label Encoder for encoding every categorical value and we used the python library "numpy" to label the dataset after converting outputs into numerical values.
2. *Feature Scaling:*
   We made scaling the data down to a range that will be easily evaluated by the algorithms and will be clearly understood for training and testing the dataset.
   The continuous attributes don't have any effect on calculations, so we used MinMaxScaler to scale the data.

C. *Training and testing dataset:*
We split the dataset into two groups:

1. Training Data.
2. Testing Data.

This is just to fit our model into the training data and make predictions based on the testing data. After doing all the previous steps, we want to avoid both over-fitting and under-fitting because they both will affect accuracy.
In the case of over fitting, the model will have high accuracy on training data but will have low accuracy using new inputs. This scenario occurs when the model is not generalized which means that the outcomes can be generalized by anyone and from here cannot make any inferences on every other data.
While the model is under-fitting when it doesn't fit into the training data so; we cannot generalize it to new inputs.

To avoid both the scenarios we can use Train/test-split. By using the scikit-learn library and the train_test_split method we can split our dataset into desired ratios. test_size=0.2 inside the train_test_split function means that the split ratio is 80:20, which means 80% of data is assigned for training the model and the rest 20% is assigned for testing the model.

We will train and test the dataset using the following seven classification algorithms:

1. Logistic regression
2. Random forest
3. Support vector machine
4. Naïve Bayes
5. Discriminant analysis
6. Decision tree
7. K nearest neighbour

D. *Feature Selection:*
The selection of attributes is very important because in the case of including redundant features it can lead to inaccurate predictions. The existence of many categorical attributes turns the feature selection process into a difficult task.

To select the attributes we used a model selection from sklearn, but every time the selected attributes were not the same as the previous set of selected attributes so that the best-selected set of features could be determined after comparing the importance indices from (logistic regression – random forest – naïve Bayes- decision tree) :

 

TABLE1: IMPORTANCE INDICES OF FEATURES USING THE MODEL SELECTION FROM SKLEARN

| **Feature /Algorithm** | **Random forest** | **Logistic regression** | **Naïve bayes** | **Decision tree** |
|---|---|---|---|---|
| **Address** | False | False | False | False |
| **Activity** | True | False | True | True |
| **Gender** | False | False | False | False |
| **Marital** | False | False | True | False |
| **Age** | True | False | False | False |
| **Income** | True | True | False | True |
| **Loan** | True | True | False | True |
| **Duration** | False | True | False | False |
| **Type** | False | True | True | False |
| **I-score** | False | False | False | False |

According to the previous table, the selected set of features after comparing importance indices from the previous algorithms will be:

1. Activity
2. Income
3. Loan

## V. Experimental Results

We can use several performance measurements from the sklearn library to calculate the performance of the classification algorithms:-

A. *Precision*: the number of correct positive divided by the number of all positive results. The precision inputs are the actual labels and the predicted labels.

TABLE 2: PRECISION RESULTS BEFORE FEATURE SELECTION

| **Algorithm** | **Precision** |
|---|---|
| **Logistic Regression** | **86.49%** |
| **Random Forest** | **95.12%** |
| **Support Vector Machine** | **86.05%** |
| **Naïve Bayes** | **86.72%** |
| **Discriminant Analysis** | **88.89%** |
| **Decision Tree** | **94.99%** |
| **K-Nearest Neighbor** | **94.02%** |

TABLE 3: PRECISION RESULTS AFTER FEATURE SELECTION

| **Algorithm** | **Precision** |
|---|---|
| **Logistic Regression** | **86.74%** |
| **Random Forest** | **96,68%** |
| **Support Vector Machine** | **86.87%** |
| **Naïve Bayes** | **86,76%** |
| **Discriminant Analysis** | **86,76%** |
| **Decision Tree** | **96,70%** |
| **K-Nearest Neighbor** | **95,60%** |

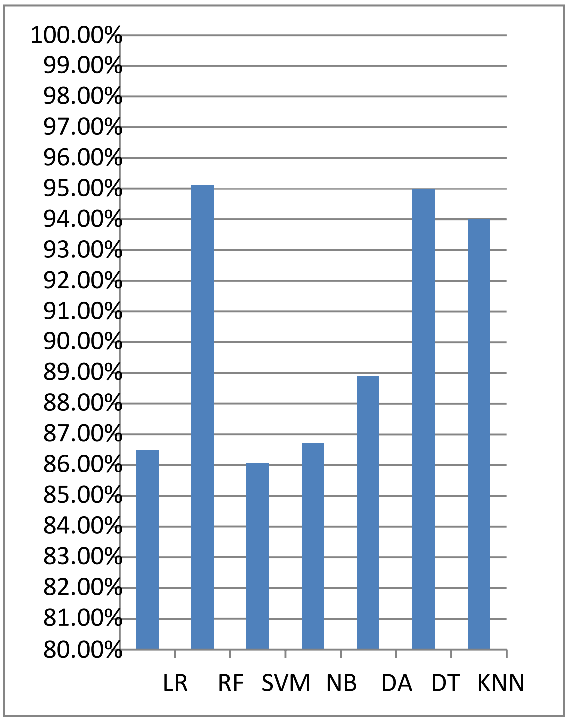


Fig1: Precision before feature selection

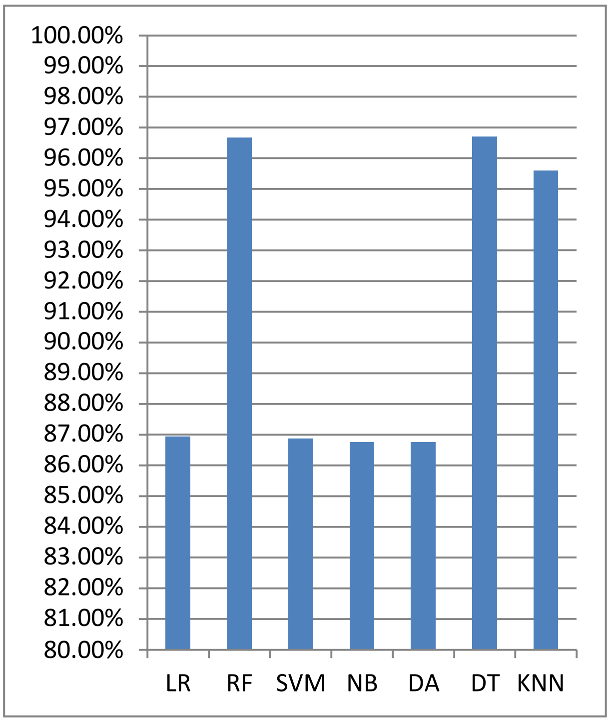


Fig2: Precision after feature selection

B. *Recall*: the number of correct positive results divided by all samples should have been classified as positive. The recall inputs are the actual labels and the predicted labels.

TABLE4: RECALL BEFORE FEATURE SELECTION

| Algorithm | Recall |
|---|---|
| **Logistic Regression** | **94.69%** |
| **Random Forest** | **95.49%** |
| **Support Vector Machine** | **94.52%** |
| **Naïve Bayes** | **94.51%** |
| **Discriminant Analysis** | **92.06%** |
| **Decision Tree** | **93.74%** |
| **K-Nearest Neighbor** | **95.34%** |

TABLE5: RECALL AFTER FEATURE SELECTION

| Algorithm | Recall |
|---|---|
| **Logistic Regression** | **94,45%** |
| **Random Forest** | **96,80%** |
| **Support Vector Machine** | **99,99%** |
| **Naïve Bayes** | **94,45%** |
| **Discriminant Analysis** | **94,45%** |
| **Decision Tree** | **96,81%** |
| **K-Nearest Neighbor** | **97,05%** |

 

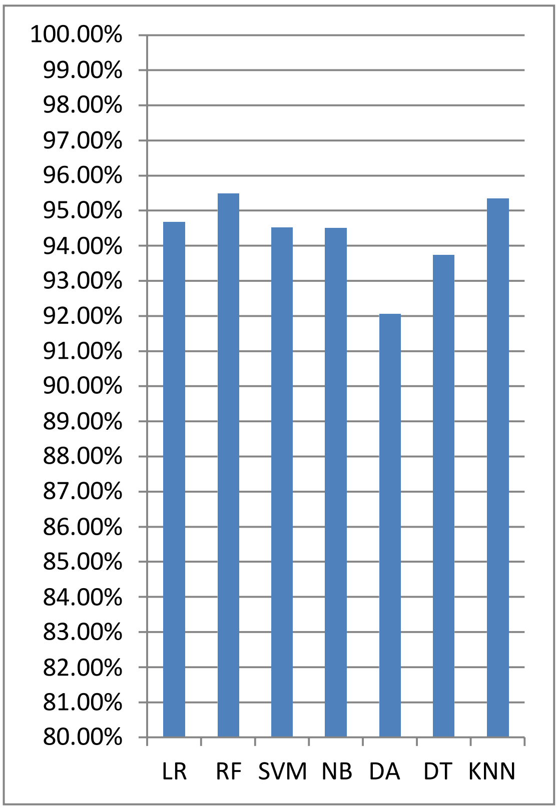


Fig 3: recall before feature selection

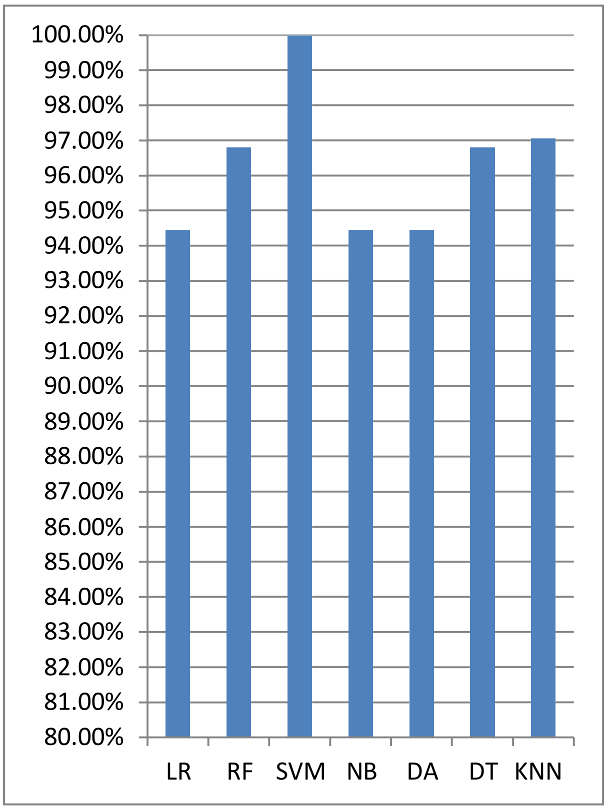


Fig 4: recall after feature selection

C. *Classification Accuracy*: the ratio of number of correct results to the total number of samples. The accuracy score inputs are the actual labels and the predicted labels.

TABLE6: ACCURACY BEFORE FEATURE SELECTION

| **Algorithm** | **Accuracy** |
|---|---|
| **Logistic Regression** | **84.08%** |
| **Random Forest** | **92.54%** |
| **Support Vector Machine** | **83.52%** |
| **Naïve Bayes** | **84.19%** |
| **Discriminant Analysis** | **84.59%** |
| **Decision Tree** | **91.12%** |
| **K-Nearest Neighbour** | **91.51%** |

TABLE7: ACCURACY AFTER FEATURE SELECTION

| **Algorithm** | **accuracy** |
|---|---|
| **Logistic Regression** | **84,16%** |
| **Random Forest** | **94,83%** |
| **Support Vector Machine** | **88,02%** |
| **Naïve Bayes** | **84,18%** |
| **Discriminant Analysis** | **84,18%** |
| **Decision Tree** | **94,85%** |
| **K-Nearest Neighbour** | **94,13%** |

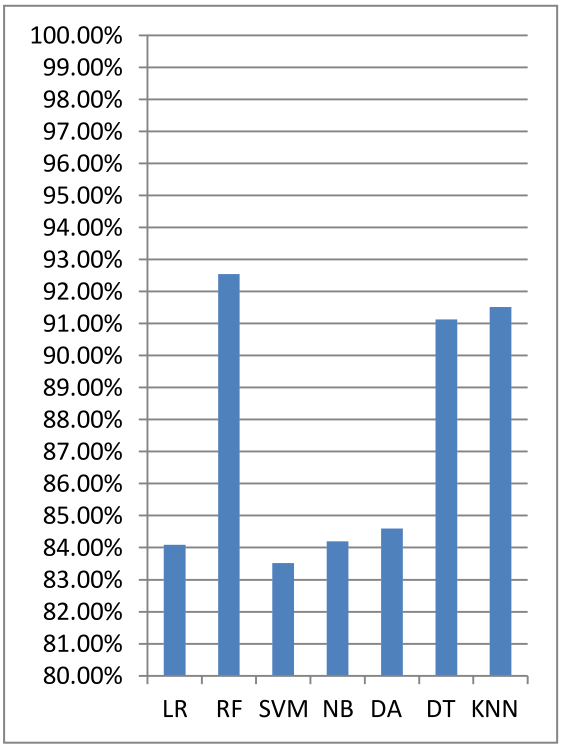


Fig 5: Accuracy before feature selection

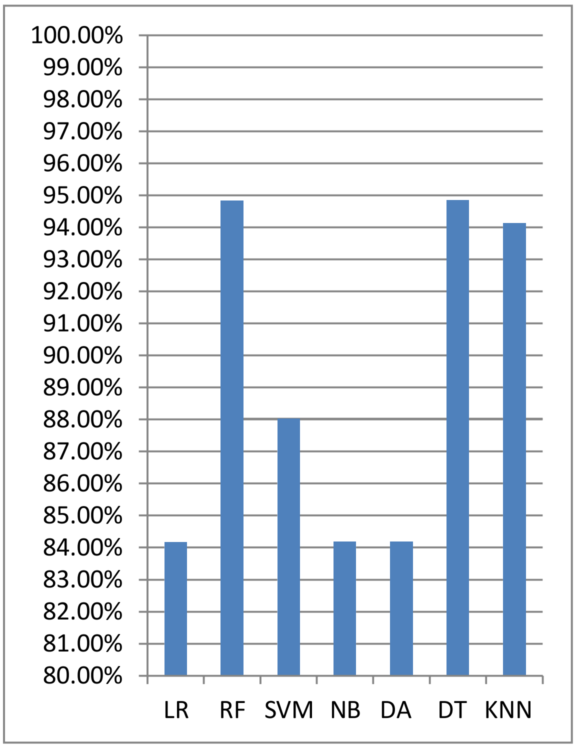


Fig 6: Accuracy after feature selection

D. *F1-score*: conveys the balance between recall and precision.it is computed as:-

2*((precision * recall) / (precision + recall))

We can get the F1- score from sklearn, as its inputs are the actual labels and the predicted labels.

TABLE8: F1-SCORE BEFORE FEATURE SELECTION

| **Algorithm** | **F1-score** |
|---|---|
| **Logistic Regression** | **90.40%** |
| **Random Forest** | **95.30%** |
| **Support Vector Machine** | **90.09%** |
| **Naïve Bayes** | **90.45%** |
| **Discriminant Analysis** | **90.45%** |
| **Decision Tree** | **94.36%** |
| **K-Nearest Neighbor** | **94.68%** |

TABLE9: F1-SCORE AFTER FEATURE SELECTION

| **Algorithm** | **F1-score** |
|---|---|
| **Logistic Regression** | **90.43%** |
| **Random Forest** | **96.74%** |
| **Support Vector Machine** | **92.97%** |
| **Naïve Bayes** | **90.44%** |
| **Discriminant Analysis** | **90.44%** |
| **Decision Tree** | **96.75%** |
| **K-Nearest Neighbor** | **96.32%** |

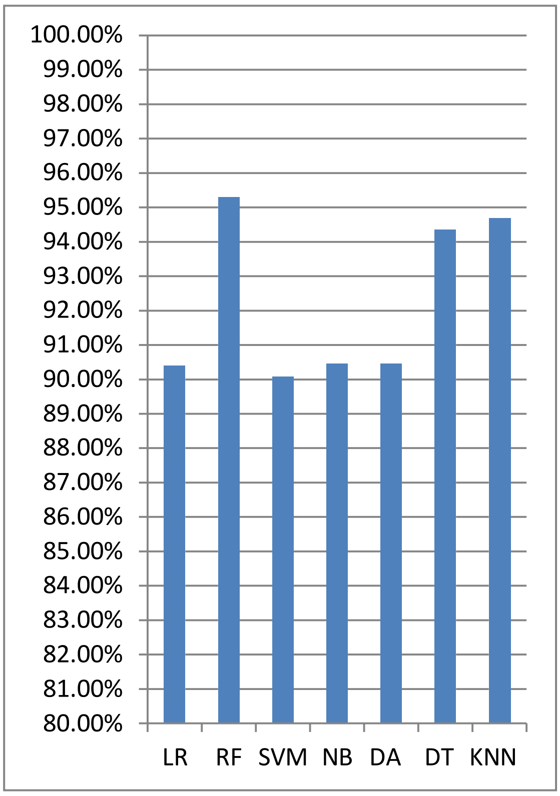


Fig 7: f1-score before feature selection

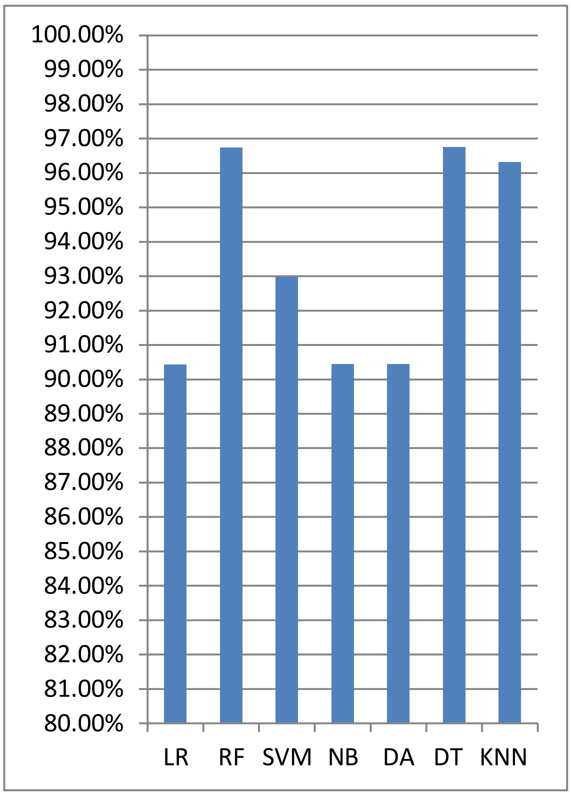


Fig 8: f1-score after feature selection

We can summarize results of (Precision - Recall - Accuracy - F1-score) before and after feature selection in the following two tables:-

TABLE10: RESULTS BEFORE FEATURE SELECTION

| Algorithm | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| **LR** | 86.49% | 94.69 | 90.40% | 84.08% |
| **RF** | 95.12% | 95.49 | 95.30% | 92.54% |
| **SVM** | 86.05% | 94.52 | 90.09% | 83.52% |
| **NB** | 86.72% | 94.51 | 90.45% | 84.19% |
| **DA** | 88.89% | 92.06 | 90.45% | 84.59% |
| **DT** | 94.99% | 93.74 | 94.36% | 91.12% |
| **KNN** | 94.02% | 95.34 | 94.68% | 91.51% |

TABLE11: RESULTS AFTER FEATURE SELECTION

| Algorithm | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| **LR** | 86.74% | 94,45% | 90.43% | 84,16% |
| **RF** | 96,68% | 96,80% | 96.74% | 94,83% |
| **SVM** | 86.87% | 99,99% | 92.97% | 88,02% |
| **NB** | 86,76% | 94,45% | 90.44% | 84,18% |
| **DA** | 86,76% | 94,45% | 90.44% | 84,18% |
| **DT** | 96,70% | 96,81% | 96.75% | 94,85% |
| **KNN** | 95,60% | 97,05% | 96.32% | 94,13% |

 

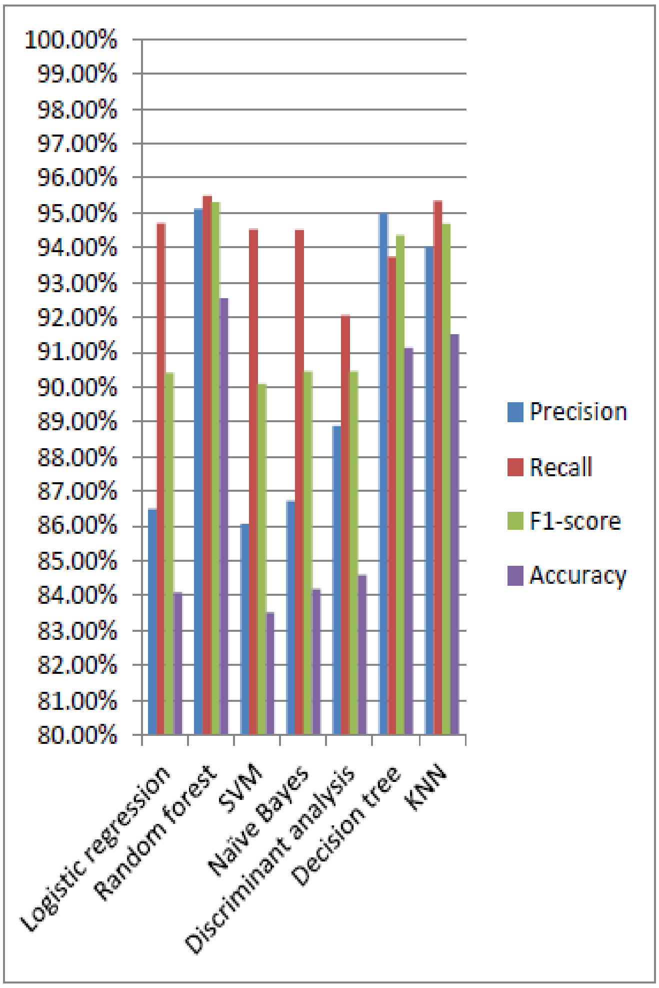


Fig 9: results before feature selection

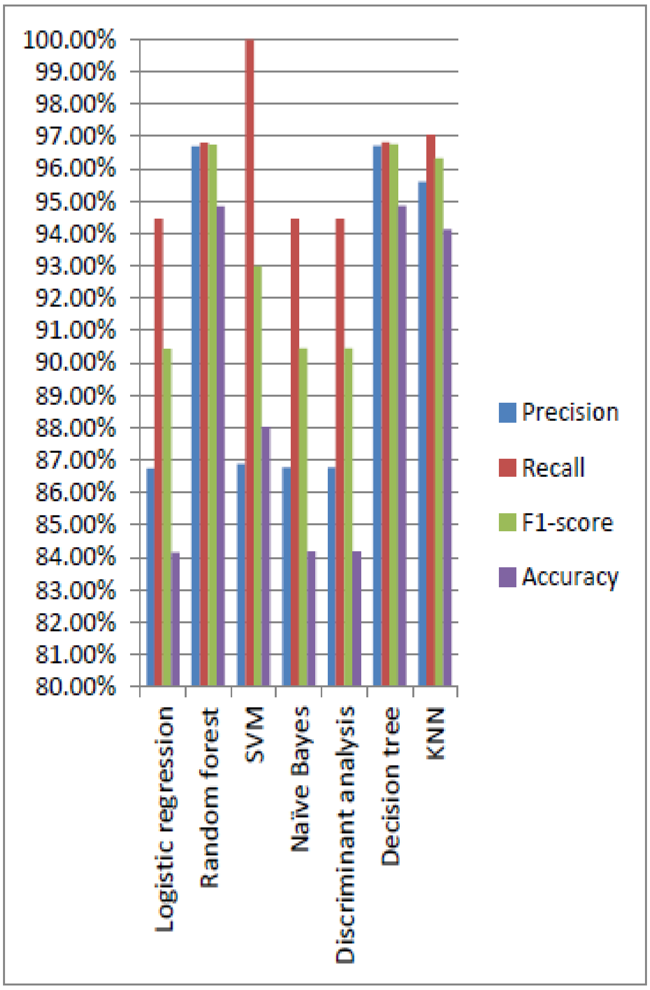


Fig 10: results after feature selection

## VI. CONCLUSION

This paper is considered a step towards an efficient prediction approach for default customers of personal loans using machine learning techniques. Through this paper, we have attempted to get the best approach to predict Credit Defaults. A description of the ABE personal loans dataset is presented as a basis for evaluation. We found that we can use the most important features (activity – income – loan) after comparing the importance indices of four algorithms that can lead to high accuracy for the prediction of default customers out of ten features. We have observed that the Decision tree has the best performance than any other machine learning algorithm for our dataset with significant prediction accuracy of almost 94.85% although, Random forest and KNN also performed well but with lower accuracy than the Decision tree.

# REFERENCES

**[1].** H. Abdou and J. Pointon, "Credit scoring and decision making in Egyptian public sector banks", International Journal of Managerial Finance, vol. 5, no. 4, pp. 391-406, 2009.
**[2].** J. H. Bank and G. Davies, History of Money: From Ancient Times to the Present Day. University of Wales, 2002, pp.48-53.
**[3].** B. Clarke, E. Fokoue, and H. H. Zhang, Principles and theory for data mining and machine learning, Springer Science & Business Media, 2009.
**[4].** P. H. Collin, Dictionary of banking and finance. London: A & C Black, 2005.

 

**[5].** K. Das and R. N. Behera, "A survey on machine learning: concept, algorithms and applications," International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, pp. 1301–1309, 2017.

**[6].** A. Dey, "Machine learning algorithms: a review," International Journal of Computer Science and Information Technologies, vol. 7, pp. 1174–1179, 2016.

**[7].** D.Graeber , Debt: The first five thousand years, Melvillehouse Brooklyn, New york ,2011.

**[8].** K. Hasan, "Development of a credit scoring model for retail loan granting financial institutions from frontier markets," International Journal of Business and Economics Research, vol. 5, pp. 135–142, 2016.

**[9].** R. Edwards and U. Rothbard, "Commercial Banks &," Historical Development." Journal of Applied Sciences Research, vol. 7, no. 7, pp. 1024–1029, 2000.

**[10].** S. Imtiaz and A. J., "A better comparison summary of credit scoring classification," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 7, 2017.

**[11].** H. Ince and B. Aktan, "A comparison of data mining techniques for credit scoring in banking : A managerial perspective", Journal of Business Economics and Management, vol. 10, no. 3, pp. 233-240, 2009.

**[12].** A. Jafar Hamid and T. Ahmed, "Developing Prediction Model of Loan Risk in Banks Using Data Mining", Machine Learning and Applications: An International Journal, vol. 3, no. 1, pp. 1-9, 2016.

**[13].** K. Abbas and N. Yousefi, "A proposed classification of data mining techniques in credit scoring.",International Conference of industrial Engineering and Operations Management, Kuala Lumpur, Malaysia, 2011, pp. 22-4.

**[14].** K. Raj and R. Verma,"Classification algorithms for data mining: A survey", International Journal of Innovations in Engineering and Technology (IJIET), vol. 1, no. 2, pp. 7-14, 2012.

**[15].** S. Mahdi, M. Gholamian, M. Siami and Z. Hajimohammadi, "Credit scoring in banks and financial institutions via data mining techniques: A literature review", Journal of AI and Data Mining, vol. 1, no. 2, pp. 119-129, 2013.

**[16].** M. de M. Sousa, Instituto Federal Goiano Campus Ceres, GO, Brazil, R. S. Figueiredo, and Federal University of Goiás, GO, Brazil, "Credit analysis using data mining: Application in the case of a credit union," J. Inf. Syst. Technol. Manag., vol. 11, no. 2, pp. 379–396, 2014.

**[17].** S. Srivastava and G. A. R. G. Anchal, "Data mining for credit card risk analysis," a Review. "International Journal of Computer Science, vol. 3, no. 2, pp. 193–200, 2013.

**[18].** M. Sudhakar and D. C. V. K. Reddy, "Credit evaluation model of loan proposals for banks using data mining," International Journal of Latest Research in Science and Technology, pp. 126–131, 2014.

**[19].** L. C. Thomas, "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers," Int. J. Forecast., vol. 16, no. 2, pp. 149– 172, 2000.

**[20].** W. Xiao, Q. Zhao, and Q. Fei, "A comparative study of data mining methods in consumer loans credit scoring management," J. Syst. Sci. Syst. Eng., vol. 15, no. 4, pp. 419–435, 2006.

**[21].** I.-C. Yeh and C.-H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Syst. Appl., vol. 36, no. 2, pp. 2473–2480, 2009.

 