FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 6 No. 3, June, 2022, pp 41 - 50
DOI: https://doi.org/10.33003/fjs-2022-0603-830

LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING APPROACH

***Nureni Ayofe Azeez and Adekola Oluwatoniloba Emmanuel**

Department of Computer Sciences, Faculty of Science, University of Lagos, Lagos, Nigeria.

*Corresponding authors' email: nurayhn1@gmail.com, atoniloba@gmail.com

ABSTRACT

Banks have various goods to sell in the banking system. The major source of income and profit, however, is their credit lines. As a result, they can profit from the interest on the loans they credit. The profit or loss of a bank is mostly determined by loans, that is, whether consumers repay the loan or default. The bank can lower its Non-Performing Assets by forecasting loan defaulters. Previous research in this age has revealed that there are numerous techniques for studying the subject of loan default control. However, because accurate forecasts are critical for profit maximization, it is critical to investigate the nature of the various methodologies and compare them. In this research, the datasets used were gathered from Kaggle for training and testing. The results gotten from both datasets were compared to ascertain which algorithm could best be used for predicting loan approval and also to determine which features are most important in predicting loan approval. The different metrics of performance that were used to define the results are: Accuracy, Precision, Recall and F1-Score. Eight different algorithms were used to train the models, these are: the Logistic Regression algorithm, Random forest, Decision trees, Linear Regression, Support Vector Machine (SVM), Naïve Bayes, K-means and K Nearest Neighbors (KNN) algorithms. The final results revealed that the models generated varied outcomes. From the results shown across both datasets, Logistic regression had - 83.24% and 78.13% of accuracy, followed by Naïve Bayes with 82.16% and 77.34% accuracy level, Random Forest garnered 81.08% and 78.91% level of accuracies for both dataset and Linear Regression had 80.35% and 78.56%. Linear Regression had precision of 84.21% and 79.25%, followed by logistic regression - 82.39% and 78.98% and Naïve Bayes with 83.35% and 76.70%. Finally, Logistic regression had 97.76% and 78.13% level of sensitivities while Support vector machine is considered to be least sensitive with - 72.24% and 66.41%.

Keywords: *Curcuma longa* powder, metal concentration, Health risk assessment

INTRODUCTION

A loan is essentially an arrangement between two parties (the lender and the borrower) in which the lender gives the borrower a loan (money, property, or other tangible commodities) based on the lender's conviction in the borrower's ability to repay the loan plus interest at a later date. In the case of personal loans, there is no criteria by which the lender can determine if the borrower is capable of repaying the loan plus interest on time.

Almost every bank's principal operation nowadays involves the approval and distribution of loans. Profits from loans distributed by the bank account for a significant component of a bank's assets. The primary goal in a banking setting is to place assets in trustworthy hands where a guaranteed profit is highly likely.

A rigorous loan approval process is now used by many financial organizations and banks, but there are no guarantees that the applicant picked is deserving of the loan or that he represents the lowest risk of defaulting on the loan and being able to totally pay back the loan when it is d8ue. To anticipate whether or not a loan will be provided to a particular applicant, this system uses machine learning approaches to validate feature sets automatically. Because it places varied emphasis on different factors, previous models have a drawback: in reality, loans are occasionally accepted solely on the basis of one strong factor, which is impossible with this method.

Employees of financial institutions such as banks, money lending firms, and others, as well as borrowers submitting loan applications, will find Loan Prediction extremely useful. This study seeks to give a faster, more efficient, and simpler method of sorting through loan applicants and selecting deserving individuals with a high likelihood of repaying the loan. It may provide the bank with unique benefits.

In the lending/banking industry, the two most important questions are:

a) Is the borrower a high-risk borrower?
b) Should the bank lend to the borrower, given his/her risk?

The borrower's interest rate is determined by the answer to the first question. The riskiness of the borrower is measured by the interest rate, which is determined by factors such as time value of money. The higher the interest rate, the riskier the borrower. Bank can then decide if the applicant is eligible for the loan based on the interest rate. Borrowers receive loans from investors (lenders) in exchange for the prospect of interest-bearing payback. That is, the lender only earns money (interest) if the borrower repays the loan. The lender, on the other hand, loses money if he or she does not repay the loan.

Automatic weighting of loan processing criteria will be done by the Loan Prediction System, and new test data will be processed with the same weightings. The applicant can be given a deadline to determine whether or not his or her loan will be approved. The Loan Prediction System allows you to jump to a single application and check it depending on its priority. This paper is intended solely for the Bank/Finance Company's controlling authority; the entire prediction process is conducted in private, and no stakeholders will be able to influence the outcome. The results for a specific Loan Id can be communicated to the bank's relevant departments, who can then take necessary action on the application. This facilitates the completion of the remaining

requirements by the other department. This paper focuses on building a platform which is based on information provided by the borrower. It calculates if the borrower is qualified to take such a loan and predicts the likelihood of the borrower paying back such loan.

It is a win-win situation for everyone when it comes to loan applicants and bank workers. It is the goal of this paper to provide a fast, simple, and effective way to choose competent applicants for employment. In the long run, it could provide the bank with distinct advantages. It is possible for the Loan Prediction System to automatically calculate the weight of each feature involved in loan processing, and the same characteristics are applied to new test data in line with the weights assigned to them. Whether or not a loan application will be granted can be determined by setting a deadline for the applicant to meet. In order to quickly assess a specific loan application, any bank can adopt this Loan Prediction System.

For a long time, patterns in loan default have been examined from a socioeconomic perspective. For the most part, economists believe that empirical modeling of these complex systems is necessary to estimate the likelihood that a borrower will fail on a loan. The application of machine learning for such operations is now a trend that is prominent. Learning, like intellect, encompasses a vast range of processes that are difficult to pinpoint. Adaptations to systems that execute AI-related activities are referred to as machine learning. Recognition, diagnosis, planning, robot control, and prediction are a few examples of these types of activities. Any improvements to current systems or novel combinations of the existing systems may be included in these revisions (Nilsson, 1998). Prior to producing predictions, the machine learning tasks can be applied on a sample of test data. Training is the term used to describe this process. Sample data is usually 70% trained, with the remaining 30% being used to assess how accurate the machine learning task can predict outcomes (Kumar et al., 2019).

In 1943, scientist Warren McCulloch, a neurophysiologist, and mathematician Walter Pitts published a study on neurons and how they work, which was the first case of neural networks. The first neural network was born from their choice to develop a model based on this utilizing an electrical circuit (Mayo et al., 2018).

Loan default prediction and credit rating have become increasingly important in recent years as consumer financing has skyrocketed. Credit scoring employs a variety of statistical models, including logistic regression, Naive Bayes, probit analysis, and liner discriminant analysis. When dealing with non-linear relationships, however, these strategies tend to perform badly. As a result, matching a particular statistical assumption in a practical application is rather tough. Machine learning approaches applied to credit scoring, on the other hand, have been found to produce greater outcomes and accuracy than statistical analysis. These methods include: Artificial Neural Networks, Support vector machine (SVM), Random forest, Decision trees etc. (Niu et al., 2019).

There are two sorts of credit scoring systems: fresh credit application verdicts and loan default prediction after lending. The first type calculates a credit score based on the loan candidate's personal information and financial situation; a higher score indicates a higher possibility of loan acceptance, while a low score indicates a larger risk for the loan. The second category is concerned with the loan applicant's credit history. Based on a review of an applicant's credit history, a financial institution can estimate the loan amount that an application is likely to safely repay (Zhao et al., 2015).

In an attempt to select and identify genuine loan applicant, Gomathy et al., in 2021 developed a system through which this could be achieved. The authors achieved this with the aid of machine learning technique. The system permits an eligible applicants to be automatically selected based on the available criteria. The prediction was attained with the aid of Decision Tree algorithm. The main challenge with this approach is the adoption of only one machine learning algorithm (Gomathy et al., 2021).

Sheikh et al. adopted a machine learning technique for predicting loan defaulters. The Logistic regression model was used along the dataset obtained from Kaggle for prediction. The results obtained were compared to ascertain its effectiveness by using specificity and sensitivity as metrics. The final results show a greater performance with similar projects. It is, however, noted that the results are not a true representation of the expected results as only one machine learning algorithm was used without any justifiable reason (Sheikh et al., 2020)

Having realized the importance of loan prediction in the current day banking system, Sujatha et al. developed a web-based application to carry out extensive and more reliable prediction using logistic regression which was implemented in Python programming language. The system can deliver high accuracy results and moderate loss for training and validate data. It is however, noted that the system's performance is limited with certain features and cannot assist the users beyond those limits (Sujatha et al., 2021).

Arun et al. adopted six machine learning algorithms for prediction of android applications with the intention of reducing this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. The objectives of their work was achieved by mining the profiles of those who have been offered similar loan in the past. The result of this work efficient when compared to similar work. The profiles used in getting those eligible for loan are very limited hence could not represent the entire population (Arun et al., 2016).

Foster et al. has undoubtedly contributed to the same research by carrying out loan defaults and hazard models for bankruptcy prediction. The authors examined if results vary when loan default status and/or audit opinion parameters are excluded from hazard bankruptcy prediction models. The reseatrch uses logistic regression to identify variables for parsimonious bankruptcy prediction models to validate hypotheses. The results improve the accuracy for financially challenged samples with hazard model (Foster et al., 2013).

Ravisankar et al. adopted data mining approach to identify industries that recourse to financial statement fraud. Six algorithms were used and tested on 202 Chinese companies' dataset and compared without and with feature selections. Of all the six data mining techniques, Probabilistic Neural Network (PNN) had the best performance when there was no feature selection. With feature selection, GP and PNN outperformed others with nearly the same level of accuracies (Ravisankar et al., 2011).

Bekhet and Eletter in 2012 attempted to develop a model with Artificial Neural Networks (ANN) as a decision support system for Jordanian commercial banks to assist in credit approval evaluation. The system can be easily utilized by credit officers in taking excellent decisions before determining future loan

applications and applicants. The results obtained with this technique showed that ANN was better in performance. The main challenge with the approach was in the small dataset used for the evaluation (Bekhet and Eletter, 2012).

Both manual loans approval as well as traditional algorithms approaches have been characterized with low performance and recognition rate. Against this backdrop, Zhang and Li in 2018 proposed an integrated learning classification model that utilized Particle swarm optimization (PSO) optimization support vector machine (SVM). PSO was used to optimise SVM while AdaBoost was used to integrate SVM weak classifier while prediction model was established. It was observed that the AdaBoost-PSO-SVM approach can successfully enhance the level of accuracy (Zhang and Li in 2018).The major challenge is the small number of sample used for the classification.

An enhanced model of machine learning technique was developed to determine the credit or loan worthiness of bank customer in Nigeria. In an attempt to test the effectiveness of this model, a very reliable and applicable dataset was obtained from UCL repository. The efficiency of the model was shown with the aid of confusion matrix with accuracy as the major metric (Asogwa, 2019).

Łuczak et. al. attempted to compare the performance of seven classifiers with the intention of establishing credit worthiness of potential clients. They applied two datasets. The challenge of the work is in the inability to conduct ensemble for better performance and evaluation (Łuczak et al., 2021).

Machine Learning for a Loan Approval System

The following are the machine learning algorithms applied in this work: Decision Trees (Zhu et al., 2019), Logistic Regression (Vaidya, 2017), (Fenjiro, 2018), Support Vector Machines (SVM) (Goyal and Kaur, 2016), Naïve Bayes (Hamid and Ahmed, 2016), Random Forests (Soni and Varghese, 2019), K-Means, K-Nearest Neighbors (KNN) (Sudhamathy, 2016), and Linear Regression.

Table-1 Comparing the pros and cons of various algorithms.

ALGORITHM	APPROACH	STRENGTH	WEAKNESS
Decision Trees	Divides the dataset into many branches that maximize the amount of information gained from each split to learn in a hierarchical approach by splitting it repeatedly.	• Learn non-linear connections with little difficulty • Can withstand outliers	• Over fitting is more likely with unconstrained trees because they can keep branching until the training data is memorized.
Logistic Regression	Decision trees' categorization counterpart. The logistic function maps predictions from 0 to 1, thus they can be thought of as class probabilities.	• The probabilistic interpretation of the outputs is quite nice. • It is possible to regularize the algorithm such that it does not over fit the data. • New data can be readily added.	• When faced with various or non-linear choice limits, this trait tends to falter. • Complex relationships cannot be captured because the system is not flexible enough.
Support Vector Machines (SVM)	Utilizes a process known as kernels, which works by calculating the distance between two points. A decision boundary is then discovered using the method, which optimizes the distance between the nearest members of different classes between one another.	• There are a large number of kernels from which to select. • Non-linear decision boundaries can be modeled using this method. • Even in high-dimension spaces, it's fairly strong against over fitting.	• High demands on the computer's memory • Due of the necessity of selecting the proper kernel, tuning might be more difficult. • Additionally, it does not scale well when working with larger datasets.
Naïve Bayes	Using conditional probability and counting, this algorithm is incredibly easy. In essence, the model is a table of probabilities that is continually updated as new training data is collected.	• Simple to implement • Scalable with the dataset. • Simple to implement Conditional independence is rarely true, but its models nonetheless perform quite well in reality.	• Models trained using other algorithms frequently outperform it due to its simplicity.
Random Forests	Is an ensemble method that can be compared to a closest neighbor predictor in terms of effectiveness?	• The program runs quickly • Can deal with data that is imbalanced or missing, among other things.	• Can over fit noisy datasets when used for regression. • Cannot predict beyond the range of the training data when used for forecasting.

K-Means	Attempts to establish clusters of comparable objects by grouping them together. It looks for similarities between the items and then organizes them into clusters based on those similarities. Using K-means clustering, there are three steps to the process. These three steps can be broken down as follows. o The first step is to decide on the k values. o Set the centroids to their default values. o Find the average for the chosen group.	• Using K-Means instead of hierarchical clustering can save time if the variables are large. • When the clusters are spherical, K-Means clusters are more compact than hierarchical clusters.	• Unpredictability of K value • Global clusters are not compatible with it. • Depending on the initial partition, different final clusters may be formed. • Clusters of varying size and density do not perform well with this method.
K-Nearest Neighbors (KNN)	A simple algorithm that keeps track of all the cases and assigns a similarity score to fresh data or cases. Data points are typically categorized based on how their neighbors are classed using this technique.	• Informally, KNN goes by the moniker Lazy Learner (Instance based learning). During the training phase, it doesn't learn anything. A discriminative function is not derived from the practice data. In other words, it doesn't have a learning curve. Only while creating in-the-moment predictions does it use the training dataset stored in the system. In comparison to other algorithms that require some form of training, the KNN algorithm is light years ahead in terms of speed. • Since there is no need to train the KNN algorithm before using it to make predictions, new data can be added without affecting the system's accuracy. • KNN is a breeze to use. To use KNN, you'll need to know the value of K and some distance information (e.g. Euclidean or Manhattan etc.)	• Because calculating the distance between each new point and each existing point is so expensive in large datasets, the algorithm's performance suffers. • It is difficult for the KNN method to determine the distance in each dimension when dealing with enormous amounts of dimensional data. • Feature scaling (normalization and standardization) is required before to using the KNN method on any dataset. KNN may produce incorrect predictions if we don't do this. • As a result of this sensitivity, KNN works well with noisy data. Missing values must be manually imputed, and outliers must be removed.
Linear Regression	Summary and examination of the relationships between two continuous (quantitative) variables is possible.	• Linear regression excels when dealing with data that can be separated along a single axis. • Simpler to implement, interpret, and train on • Dimensionally reduced methods, regularization, and cross-validation are used to handle over fitting well. • In addition, extrapolation beyond a single data set has its advantages.	• The assumption of linearity between the dependent and independent variables. • It's frequently prone to generating a lot of noise and being too tight. • Outliers have a big impact on the results of a linear regression.

METHODOLOGY

At this point, we discuss the developmental process involved in the implementation of a loan approval system, they include: Dataset acquisition, Data processing and Feature extraction, Implementation algorithms and Result Analysis.

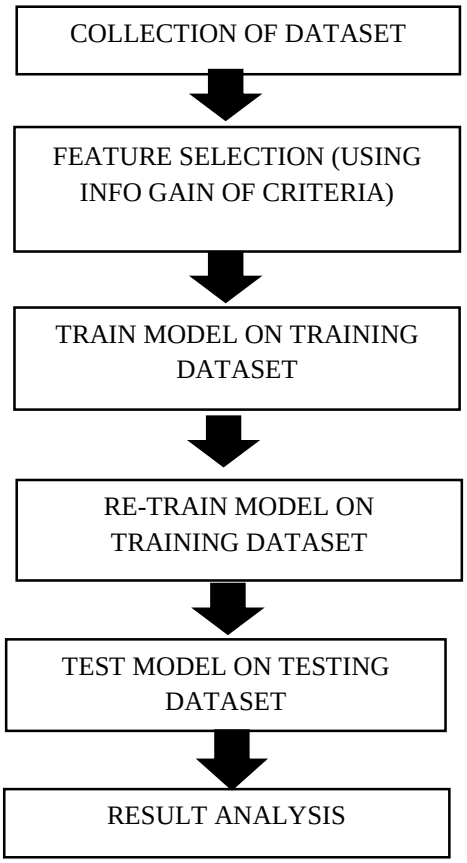


Figure 1: Loan Prediction Methodology

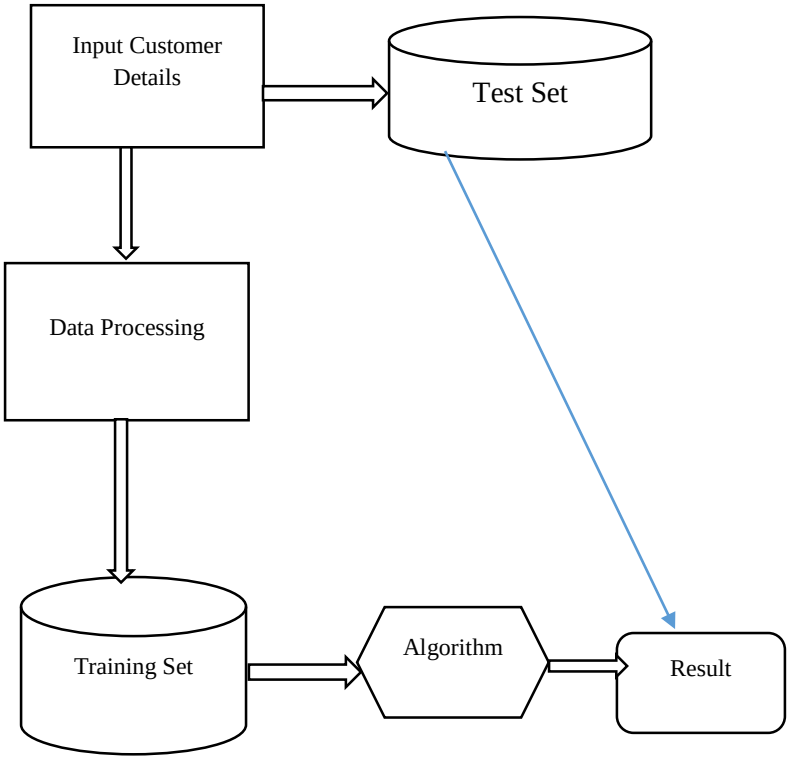


Figure 2: Training and Testing Model

Experimental Dataset Acquisition

A training dataset is required for training the selected algorithms. The local datasets used was obtained from https://www.kaggle.com/. The Loan prediction problem dataset logs of borrower statistics that were used to train and then test the algorithms for achieving the desired accurate loan prediction.

The train.csv contains logs of borrower details that was used to train the model and a total of 13 factors by which the loan was approved while test.csv contains borrower details that was used to test the models accuracy.

The various datasets are gotten from the following links are given hereunder:

1) Dataset: https://www.kaggle.com/ninzaami/loan-predication
2) Dataset: https://www.kaggle.com/kapilt1991/loan-predcition-demo/data

Data processing and Feature extraction

With the help of the learning algorithms, the train dataset was used to create a functional model, while the test dataset was utilized to assess the model's performance and predict which of the loan applicants should be approved.

Implementation algorithm

Eight algorithms were used in this work. They are:

1) **Decision trees:** As a visual and clear representation of decisions and decision making, a decision tree may be used in decision analysis. The decision-making process follows a tree-like structure, as implied by the name. In data mining, it is often used to create strategies for achieving certain objectives, but in machine learning as well. The root of a decision tree is at the top of an upside-down decision tree.

$$\textbf{Entropy} = \sum_{i=1}^{n} - Pi * \log_2(Pi) \qquad (1)$$

$$\textbf{Gini} = 1 - \sum_{i-1}^{n} (Pi)^2 \qquad (2)$$

2) **Random forest (RF):** For both classification and regression, random forest a supervised learning technique is employed. But it is mainly used for classification problems. Random forest algorithm creates decision trees on data samples and then obtains the prediction from each of them before it finally selects the best solution by means of voting. It is an ensemble type of system where we have

multiple tree (models) that are individually weak but when combined they form a power model (Azeez et al., 2021b).

3) **Logistic Regression:** Classification algorithms such as logistic regression employ supervised learning to estimate the likelihood of certain outcomes. There are only two possible classes for the dependent variable, hence it is dichotomous in nature. This means the target/dependent variable is binary in nature. Each bit of information is either coded as a 1 or a 0 based on whether it indicates yes or success.

$$P=\frac{1}{1+e^{-(B_0+B_1X_1+B_2X_2+B_3X_3+\ldots+B_nX_n)}} \quad (3)$$

4) **Support Vector Machine (SVM):** There are supervised learning models with related learning algorithms that evaluate data used in classification and regression analysis that are known as support vector machines (also known as Support Vector Networks). Classification and regression problems can be solved using the SVM algorithm, a common machine learning tool.

$$\ell(y)=\max(0,1+\max_{y\neq t} w_y x-w_t x \quad (4)$$

5) **Naive Bayes:** A supervised machine-learning technique that makes use of the Bayes' Theorem is known as a Naive Bayes Classifier. It's based on the premise that features aren't statistically related. Because input variables are assumed to be independent of one another, the theorem is based on the erroneous assumption that they are. It's a classifier with good results, regardless of what people think. It is possible to calculate posterior probability P(c|x) using the Bayes theorem by combining P(c), P(x), and P(x|c). Take a look at the following equation:

$$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)} \quad (5)$$

$$P(c\mid x)=P(x_1\mid c)*P(x_2\mid c)*\ldots*P(x_n\mid c)*P(c) \quad (6)$$

Here,

- For a given predictor, what is the posterior probability that the target class will be in? (Attribute).
- P(c) denotes the prior probability of a particular class occurring.
- There is a probability of predictor given class called P (x|c).
- The prior probability of the predictor is given by P(x).

6) **K Nearest Neighbor (KNN**): Classification and regression problems can be solved with K Nearest Neighbor. When dealing with classifications challenges, it's more typically used in the industry. There is a simple algorithm known as the K nearest neighbor's approach that keeps track of all existing instances and assigns new ones a classification based on a majority vote from those examples' k closest neighbors. A distance function established that the case that has been assigned to the class is the most common among its K nearest neighbors (Azeez et al., 2021a).

Distance functions include things like Euclidean, Manhattan, Minkowski, and Hamming distances. Three of the four arithmetic operations on continuous functions are used, whereas Hamming is used on categorical variables. If K is 1, the case will be assigned to the class of its nearest neighbor. When using kNN modeling, it can be challenging to decide on the number of turns to take (Azeez et al., 2020).

Euclidean:

$$\sqrt{\sum_{i-1}^{k}(x_i-y_i)} \quad (7)$$

Manhattan:

$$\sum_{i-1}^{k}|x_i-y_i| \quad (8)$$

Minkowski:

$$(\sum_{i-1}^{k}(|x_i-y_i)^q)^{\frac{1}{q}} \quad (9)$$

7) **K-Means:** Clustering can be solved using an unsupervised technique. It uses a series of clusters to classify a given data set in a plain and simple manner (assume k clusters). To peer groups, data points inside a cluster are homogeneous and heterogeneous.

Remember when you used to make shapes out of ink blots? This behavior is related to what k signifies. To figure out how many different clusters / populations are present, you look at the form and dispersion!

How K-means forms cluster:

- K-means selects k centroids, or points, for each cluster.
- The closest centroids of each data point form a cluster, i.e. k clusters.
- It locates each cluster's centroid using the members already present in the cluster. New centroids can be found here.
- Steps 2 and 3 must be repeated as fresh centroids are acquired. Find the distance between new centroids and each data point, and then join those new k-clusters. In other words, keep going until the centroids don't change anymore.

8) Linear Regression: It's a tool for estimating continuous-variable real values, such as house costs, phone calls, and total sales (s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation **Y= a *X + b**.

Reliving this childhood event is the best way to grasp linear regression. Let's imagine you ask a fifth-grader to arrange everybody in his class in ascending weight order without asking them their weights! What do you expect the kid to do? He or she would most likely glance at people's height and build (visually assess) and arrange them based on a combination of these observable criteria. This is real-life linear regression! The child has deduced that height and build are related to weight through a connection that looks like the one shown above.

$$Y = c + m_1X_1 + m_2X_2 + ... + m_nX_n \quad (10)$$

Where:

Y= dependent Variable or Target Variable

m = slope

c = intercept

X= Independent Variables

Metrics of Evaluation

The necessary metrics and measures have been calculated. These figures include: Accuracy, Precision, Recall and F-Score (Azeez et al., 2020).

There are four important terms that have to be noted:

- True Positives (TP): When the model accurately predicts the positive class, it is considered a true positive outcome.
- True Negatives (TN): Similarly when the model accurately predicts the negative class, it is considered a true negative outcome.
- False Positives (FP): When the model inaccurately predicts the positive class, it is considered a false positive outcome.
- False Negatives (FN): Therefore, when the model inaccurately predicts the negative class, it is considered a false negative outcome.

Accuracy: In terms of performance, the most significant metric is accuracy. This is achieved by dividing the total number of observations by the total predicted observation number. Just because a model has a high degree of accuracy doesn't mean it's appropriate. For symmetric datasets, accuracy serves as a helpful statistic because similar values are common throughout the datasets. Fake positive and negative test results are almost exactly the same (Azeez et al., 2019).

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ predictions\ made} \quad (11)$$

$$\frac{(TP)+(TN)}{(TP)+(FP)+(TN)+(FN)} \quad (12)$$

Precision: When it comes to precision, it's the ratio of accurately predicted positive results to all expected positive results. Precision answers the question of how many "safe" loan applications are genuinely safe loans for the lender in this situation.

$$\frac{(TN)}{(TN)+(FN)} \quad (13)$$

Recall: There are precisely determined and anticipated positive observations in the entire class as a percentage of all observations.

$$\frac{(TP)}{(TP)+(FN)} \quad (14)$$

F1-Score: This is the weighted average of precision and recall. To account for both possible outcomes, both potential positive and negative results are considered. Normally, F1 is more important than accuracy, and this is particularly true given the unequal distribution of students. Trying to find a balance between precision and recall is much easier using this technique. When you combine high precision with low recall, you get an incredibly precise result, but it also leaves out a big number of occurrences that are difficult to classify. The better our model performs, the higher the F1 Score becomes. It can be mathematically represented as follows:

$$\frac{2*precision*recall}{precision+recall} \quad (15)$$

Table-2: Results from Dataset 1

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	83.24324	82.38993	97.76119	89.41979
Linear Regression	80.345452	84.212323	83.345452	81.24219
Decision Tree Classifier	74.59459	74.23324	74.59459	74.65433
Random Forest Classifier	81.08108	80.14864	81.08108	79.26731
Support Vector Machine	72.24325	85.24101	72.24325	75.10242
Naïve Bayes	82.16216	83.34821	82.16216	80.10530
K - Means	63.24324	72.72435	80.59701	76.05633
K Nearest Neighbors	62.70270	72.72727	77.61194	75.09025

RESULTS

The proposed methodology was built upon a set of 13 features obtained by processing the contents of the posts in the data set comprising of 81,820 ham posts and 5,263 phish posts.

At the end of the implementation, to determine their performance with the given measures, all classifiers were ran against the two datasets. The results are shown in Tables 2 and 3.

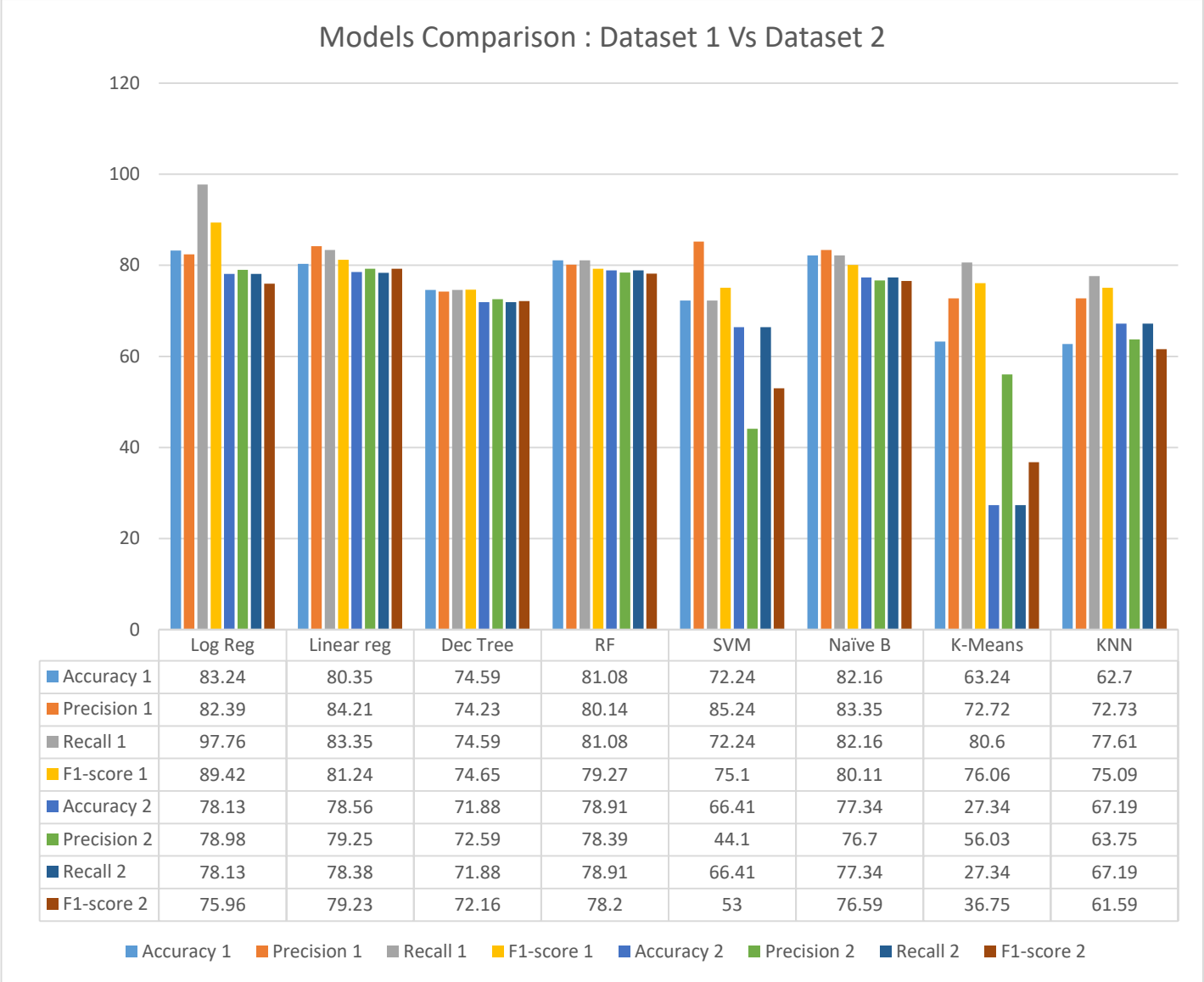
The results of the examination of metrics for classifiers using the first dataset are shown in Table 2. Accuracy, precision, F1-score, and recall are the evaluation metrics used. The classifiers are: Logistic Regression, Linear Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine, Naïve Bayes, K-Means and K-Nearest Neighbors.

From the outcome shown in Table 2, Logistic regression is the most accurate classifier, followed by Naïve Bayes, Random Forest and Linear Regression. Support Vector Machine is the most precise, followed by linear regression and Naïve bayes. Logistic regression is however the most sensitive by some margin while Support vector machine is the least sensitive

However, in Table 3 the Random Forest Classifier is most accurate with 78.906% closely followed by Linear Regression, Logistic Regression and Naïve Bayes. K-means is the least sensitive with a recall score of 27.344%, while Random Forest Classifier is the most sensitive. Linear Regression is the most precise followed by Logistic Regression and Random Forest.

Table-3: Results from Dataset 2

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	78.125	78.9790	78.125	75.96192
Linear Regression	78.5623	79.245	78.377	79.233
Decision Tree Classifier	71.875	72.588	71.875	72.161
Random Forest Classifier	78.906	78.391	78.906	78.204
Support Vector Machine	66.40625	44.09790	66.40625	53.00029
Naïve Bayes	77.34375	76.69694	77.34375	76.58925
K – Means	27.34375	56.02864	27.34375	36.74677
K Nearest Neighbors	67.1875	63.75461	67.1875	61.58731



	Log Reg	Linear reg	Dec Tree	RF	SVM	Naïve B	K-Means	KNN
Accuracy 1	83.24	80.35	74.59	81.08	72.24	82.16	63.24	62.7
Precision 1	82.39	84.21	74.23	80.14	85.24	83.35	72.72	72.73
Recall 1	97.76	83.35	74.59	81.08	72.24	82.16	80.6	77.61
F1-score 1	89.42	81.24	74.65	79.27	75.1	80.11	76.06	75.09
Accuracy 2	78.13	78.56	71.88	78.91	66.41	77.34	27.34	67.19
Precision 2	78.98	79.25	72.59	78.39	44.1	76.7	56.03	63.75
Recall 2	78.13	78.38	71.88	78.91	66.41	77.34	27.34	67.19
F1-score 2	75.96	79.23	72.16	78.2	53	76.59	36.75	61.59

Figure-3: Graphical representation of a comparison of the eight models with the metrics from dataset 1 to dataset 2.

CONCLUSION

A bank credit dataset was analyzed using a machine learning approach in this study; to predict a client's credit worthiness and loan payback capacity. Experiment were conducted with machine learning algorithms on the dataset to determine which of them was most effective for bank loan's approval. The experiment showed that all algorithms except K- Means and K-Nearest Neighbors perform brilliantly in terms of accuracy and other performance evaluation criteria. The accuracy of each of these algorithms ranges from 70% to more than 80%. In addition, efforts were made to figure out which of the elements had the biggest bearing on a customer's creditworthiness. Further research on this shall be extended to ensemble learning to enhance a dependable and reliable result.

REFERENCES

Fenjiro, Y., (2018), Machine Learning for Banking: Loan approval use case. [Online] Medium. Available at https://medium.com/@fenjiro/data-mining-for-banking-loan-approval-use-case-e7c2bc3ece3 [Accessed 20 May. 2020]

Goyal, A., & Kaur, R. (2016). Accuracy Prediction for Loan Risk Using Machine Learning Models. *Int. J. Comput. Sci. Trends Technol*, *4*(1), pp.52-57.

Goyal, A., & Kaur, R., (2016) A survey on ensemble model for loan prediction. *International Journal of Engineering Trends and Applications (IJETA)*, *3*(1), pp.32-37.

Hamid, A. J., & Ahmed, T.M. (2016). Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal (MLAIJ) Vol*. *3*(1).

Kumar, R., Jain, V., Sharma, P. S., Awasthi, S. & Jha, G. (2019) "Prediction of Loan Approval using Machine Learning", *International Journal of Advanced Science and Technology*, 28(7), pp. 455 – 460. Available at: http://sersc.org/journals/index.php/IJAST/article/view/460 (Accessed: 5, June, 2020).

Mayo, H., Punchihew, H., Emile, J. & Morrison, J., (2018) History of Machine Learning

https://www.doc.ic.ac.uk/~jce317/history-machine-learning.html viewed 28 March 2020.

Nils, N., (2021). MLBOOK – INTRODUCTION TO MACHINE LEARNING AN EARLY DRAFT OF A PROPOSED TEXTBOOK *Nils J Nilsson Robotics Laboratory Department of Computer Science | Course Hero*. [Online] Coursehero.com. Available at: <https://www.coursehero.com/file/5996903/MLBOOK/> [Accessed 24 March 2021].

Niu, B., Ren, J., & Li X. (2019) Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer-to-Peer Lending. *Information*; 10(12):397

Soni, P.M., & Paul, V. (2019). A Novel Hybrid Classification Model for the Loan Repayment Capability Prediction System. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4

Sudhamathy, G., (2016). Credit risk analysis and prediction modelling of bank loans using R. *International Journal of Engineering and Technology (IJET)*, *8*(5), pp.1-13.

Vaidya, A., (2017). Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

Zhao, Z., Xu, S., Kang, B.H., Kabir, M.M.J., Liu, Y. & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications*, *42*(7), pp.3508-3516.

Zhu, L., Qiu, D., Ergu, D., Ying, C. & Liu, K., (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*, pp.503-513.

Gomathy, C.K., Charulatha, M. S., Akash, A., & Sowjanya, M.S. (2021) The Loan Prediction Using Machine Learning" International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 10 | Oct 2021.

Sheikh, M. A., Goel A. K., & Kumar, T. (2020). An Approach for Prediction of Loan Approval using Machine Learning Algorithm," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

Sujatha, C. N., Gudipalli, A., Pushyami, B., Karthik, N., &. Sanjana, B. N. (2021). Loan Prediction Using Machine Learning and Its Deployement On Web Application," *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2021, pp. 1-7, doi: 10.1109/i-PACT52855.2021.9696448.

Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach", *IOSR J. Comput. Eng*, vol. 18, no. 3, pp. 18-21, 2016.

Foster B.P & Zurada, J. (2013). Loan defaults and hazard models for bankruptcy prediction", *Managerial Auditing Journal*, 2013.

Ravisankar, P., Ravi, V., Rao, G.R., & Bose, I (2011) "Detection of financial statement fraud and feature selection using data mining techniques", *Decision support systems*, vol. 50, no. 2, pp. 491-500, 2011.

Bekhet, H.A., & Eletter, S.F. (2012). Credit risk management for the Jordanian commercial banks: a business intelligence approach, 2012.

Zhang, T., & Li, B. (2018). Loan Prediction Model Based on AdaBoost and PSO-SVM", *International Conference on Network Communication Computer Engineering*, May 2018.

Asogwa, D.C., & Chukwuneke, C.I. (2019). Enhanced credit worthiness of bank customer in nigeria using machine learning and digital nervous system", *International educational journal of science and engineering*, vol. 2, no. 1, 2019.

Łuczak, A., Ganzha, M., & Paprzycki, M. (2021). Probability of Loan Default—Applying Data Analytics to Financial Credit Risk Prediction", *Intelligent Systems Technologies and Applications*, pp. 1-16, 2021.

Azeez, N.A., Adio, O.E., Yekinni, A.W., & Onyema, C. J. (2020) "Evaluation of Machine Learning Algorithms for Filtering and Isolating Spammed Messages" FUTA Journal of Research in
Sciences, Vol. 16(1), April, 2020: 26-38.

Azeez, N.A., Idiakose, S.O., Onyema, C.J., & Vyver C.V. (2021a) "Cyberbullying Detection in Social Networks: Artificial Intelligence Approach" Journal of Cyber Security and Mobility, Vol. 10 4, 1–30. doi: 10.13052/jcsm2245-1439.1046.

Azeez, N.A., Ihotu, A.M., & Misra S. (2021b). Adopting Automated White-List Approach for detecting Phishing Attacks" Elsevier Journal of Computers & Security 108 (2021) 102328, pp. 1-18.

Azeez, N.A., Salaudeen, B.B., Misra, S., Damasevicius, R., & Maskeliunas, R. (2019) "Identifying Phishing Attacks in Communication Networks using URL Consistency Features" International Journal of Electronic Security and Digital Forensics (InderScience). https://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijesdf