PERFORMANCE EVALUATION OF MACHINE LEARNING ALGORITHMS IN PREDICTING AGRICULTURAL LOAN DEFAULTERS

BY

OLALERE, Zainab Maruf

MTECH/SICT/2018/9196

DEPARTMENT OF COMPUTER SCIENCE
FEDERAL UNIVERSITY OF TECHNOLOGY, MINNA

NOVEMBER, 2022

PERFORMANCE EVALUATION OF MACHINE LEARNING ALGORITHMS IN PREDICTING AGRICULTURAL LOAN DEFAULTERS

BY

OLALERE, Zainab Maruf

MTECH/SICT/2018/9196

A THESIS SUBMITTED TO THE POSTGRADUATE SCHOOL FEDERAL UNIVERSITY OF TECHNOLOGY, MINNA, NIGERIA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF TECHNOLOGY IN COMPUTER SCIENCE

NOVEMBER, 2022

ABSTRACT

Financial institutions in Nigeria have continuously extended loan offers to a sector of the economy, say the manufacturing and industrial sector, compared to other sectors, like the agricultural sector, due to the peculiarity of each. To aid such underserved sectors, such as the agriculture sector, the Nigerian government has established risk-sharing interventions in the agricultural sector such as Nigeria Incentive-Based Risk Sharing System for Agricultural Lending (NIRSAL)to encourage the financial institutions towards the issuance of credits to farmers. Although the risk-sharing incentives has improvedthe issuance of loans to farmers, financial institutions still seek to reduce the leftover risk. Therefore, this research utilized a private agricultural loan dataset collected in Lavun Local Government Area of Niger state, Nigeria to predict the likelihood of agricultural loan default of farmers in Lavun Local Government Area. Dataset dimensionality reduction of Recursive Feature Elimination with Cross-validation (RFECV) and Principal Component Analysis (PCA) was appliedto the dataset to improve performance metrics. RFECV was used to reduce the features of the dataset from 60 to 44 while PCA extracted the dataset features into 31 principal components. Furthermore, machine learning algorithms of random forest, logistic regression, support vector machine, gradient boosting, and adaptive boosting were applied to the dataset. The results obtained shows that gradient boosting and random forest algorithms were the most effective when the RFECV dimension reduction technique was applied to the dataset in predicting agricultural loan defaults with precision and f1-score of 86.36% with 90.48% and 89.47% with 82.93% respectively. When PCA was applied to the dataset, logistic regression and ada boost achieved results of 78.95% and 74.35% respectively for precision and 76.92% and 74.29% respectively for f1-score. Overall, logistic regression proved to be the most consistent machine learning classifier when either PCA or RFECV is applied to the dataset while gradient boosting proved to be the best algorithm in predicting agricultural loan defaulters. The reduction of accuracy variation observed during cross-validation of the best models is proposed for further study.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

GLOSSARY OF ABBREVIATIONS

Abbreviation	**Meaning**
ABP	Anchor Borrowers' Programme
AUC	Area Under the Receiver Operating Characteristics
CRG	Credit Risk Guarantee
GDP	Gross Domestic Product
GLM	Generalized Linear Models
NIRSAL	Nigeria Incentive-Based Risk Sharing System for Agricultural Lending
PCA	Principal Component Analysis
RFECV	Recursive Feature Elimination with Cross-validation
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background to the Study

Agriculture is the art, science, and practice of cultivating plants and raising domesticated livestock to produce food, feed, and other by-products.In 2018, agriculture contributed about 4% to the global Gross Domestic Products (GDP), thus, being a crucial sector for economic growth (Adenekan & Augustus, 2021). Consequently, agriculture has continued to boost the economies of countries especially developing countries. In Africa, a large number of countries engage in agriculture of which subsistence farming is the most practised form given the prevalent economic condition in those countries (Olanipekun *et al.*, 2019). Due to the subsistence nature of agricultural input in Africa, the output obtained from the agricultural operation cannot meet needs beyond those of the farmer. Hence, low savings and investments are low among farmers given the direct relationship between savings and investments.

In Nigeria, the agricultural sector has employed over 70% of the total workforce, positioning the sector as an instrument for economic diversity and development (Adenekan & Augustus, 2021; Evbuomwan & Okoye, 2017). In the 1960s, Nigeria's export was predominantly from the agricultural sector with each region of the country playing a vital role: the north produced groundnuts, the south-east produced oil palm, and the south-west produced cocoa. However, the discovery of oil and the post-civil war rehabilitation and reconstruction contributed to the decline in agricultural exports from Nigeria (Sulaimon, 2021). Nevertheless, the rural communities within Nigeria embraced agriculture as a means of livelihood. Niger state, in the north-central region, is one of the states whose rural communities have continued with the mass production of agricultural produce, with yam and rice dominating other crops produce. For instance, rural households in Lapai, Kontagora, and Suleja regions of Niger state have recorded an average of 19 years of farming experience (Mustapha, 2019). However, given the low input to agricultural operations, the activities of farmers in the rural areas have remained subsistent thus yielding low agricultural outputs. The need to harness the vast experience of these small-holder farmers towards improving the dwindling supply of agricultural produce due to the reduced focus given to agriculture and the burgeoning population necessitated the issuance of agricultural credits to the farmers. The agricultural credits are aimed at enabling the farmers to purchase improved seeds, fertilizer, and hire mechanized farm tools.

Over the years, issuance of credit to various stakeholders in the sectors of the economy has been performed including the agricultural sector. Issuance of credit or loan involves giving an individual or a group a stipulated amount of money to enable the individual or group to purchase goods or services and return the borrowed money with the interest accrued to it. Due to the increased reliance on oil in Nigeria, most commercial banks give loans to industries in the oil and manufacturing sectors compared to the agricultural sector (Sulaimon, 2021). The varying credit allocation by banks is attributed to the risk of income and capital loss across the agricultural, manufacturing, and oil sectors. Credit financing of the agricultural sector can translate into access to critical inputs that accelerate agricultural productivity and lower barrier of entry for new entrants in the agricultural sector. In general, credit financing of the agricultural sector will lead to an increase in agricultural output. But, ease of access to such credit is limited (Ndagi*et al.*, 2016). Since the majority of the agrarian population are smallholder farmers who dwell in rural areas according to Adenekan and Augustus (2021), obtaining credit from banks without sufficient collateral becomes an uphill task. Furthermore, the problems plaguing the agricultural sector in Nigeria – volatile commodity prices, disease outbreak, and climate change – make the sector less attractive to formal financial institutions to offer credit facilities (Sulaimon, 2021).

The problem of credit financing of smallholder farmers necessitated the creation of credit risk guarantee frameworks that employ a risk-sharing model which encourages financial institutions to lend to farmers easily. The Nigeria Incentive-Based Risk Sharing System for Agricultural Lending (NIRSAL) employs this model. Aside from rendering technical assistance to farmers, modelling businesses to leverage market dynamics that benefit the agrarian population, and providing innovative insurance of agricultural products; NIRSAL offers a Credit Risk Guarantee (CRG). The CRG is a framework designed to shield financiers and investors who issue agribusiness loans against losses in a credit transaction through a risk-sharing arrangement. The NIRSAL CRG covers the risk of default on the loan principal and accrued interest to the limit of a predetermined CRG rate. Another risk-sharing programme for smallholder farmer loans is the Anchor Borrowers' Programme (ABP). The ABP seeks to boost the production of agricultural commodities and stabilizing the supply of inputs to agro-processors by providing farm inputs (cash and labour). The programme, an initiative of the Central Bank of Nigeria (CBN), employs a risk-sharing model to encourage the participation of financial institutions by absorbing 50% of the amount in default once it has been established that all means of loan recovery have been exhausted. Hence, the participating financial institutions bear the credit risk of the remaining 50% of the loan amount (Evbuomwan & Okoye, 2017). Since financial institutions will want to bear a minimal amount of risk, the financial institutions must obtain a way to predict loan defaulters while considering the peculiarities of farmers.

In other sectors of the economy such as the telecommunications and financial sectors where lending is prevalent, various machine learning methods have been used to predict loan defaulters based on the data collected about the loan requester. Machine learning, a branch of artificial intelligence, is a data analysis method that develops analytical models which learn patterns from available data and makes decisions based on the patterns established with minimal human intervention. Due to the increasing complexity in assessing credit risk, various machine learning algorithms have been applied to minimize credit risk(Zoran, 2019). In predicting loan or credit default, machine learning algorithms of support vector machine, artificial neural network, random forest, naïve Bayes, and decision tree have been used (Al-Qerem*et al.*, 2019; Dushimimana*et al.*, 2020). However, due to the peculiarity of smallholder farmers, a low amount of research has examined calculating the credit risk of the farmers.

1.2 Statement of the Research Problem

Given that banks play important roles in the market economy, its success is hinged on how well it manages the financial assets at the bank's disposal (Datkhile*et al*., 2020). As actors from the sectors of the economy approach financial institutions for funding in form of loans, these financial institutions attempt to reduce credit risk by granting loans to well-established businesses with tangible assets for collateral while turning down the request of the, presumably, risky ones. However, this act may stifle some aspects of the economy which may lead to a reduced growth rate of the overall economy. Although some governments have established several risk-sharing frameworks to encourage lending to the underserved sectors of the economy, some financial institutions are still skeptical because they want to reduce the risk to the minimum.

According to Echebiri and Onu(2019), the agricultural sector in Nigeria is plagued with uncertainties ranging from dependence on rainfall to the lack of storage facilities for perishable farm products. Consequently, it has not enjoyed the ease of obtaining loans from financial institutions compared to other sectors (Sulaimon, 2021). Since banks will want to reduce the risk of issuing loans to farmers despite the risk-sharing schemes by the government, it is imperative that a method of predicting loan defaulters among farmers be performed. Therefore, this research focuses on predicting agricultural loan defaulters among farmers. Subsequently, financial institutions can adopt the model to reduce the risk associated with loans issued to smallholder farmers.

1.3 Aim and Objectives

This aim of this research is to predict agricultural loan defaulters among smallholder farmers in Lavun, Niger State. Lavun was chosen due to the abundance of rice-based cropping enterprise in the area given the high consumption of rice in Nigeria. The objectives of theresearch are to:

i. Perform feature selection and extraction on the private Nigerian agricultural loan dataset using recursive feature elimination with cross-validation and principal component analysis respectively.

ii. Classify the instances in the dataset based on the selected and extracted features.

iii. Evaluate the performance of the classifierson feature selection and dimensionality reduction using accuracy, precision, recall, and f1-score.

1.4 Scope of the Study

This research focuses on applying machine learning approaches in predicting agricultural loan defaulters among small-holder farmers in Lavun local government area of Niger state.Feature selection and dimensionality reduction were applied to the dataset separately and their effects on machine learning performance evaluated.Other local governments within or outside Niger state were not considered.

1.5 Significance of the Study

The importance of this research cuts across financial institutions that provide loans or credits to smallholder farmers. Unarguably, financial institutions endeavour to minimize risk when issuing loans to individuals therefore, this research will guide the credit risk management framework of agricultural inclined financial institutions.

1.6 Thesis Organisation

This thesis consists of five chapters ranging from Chapter one to Chapter five. Chapter one provides an outline of the research, thestatement of the research problem, the aim and objectives of the research, the scopeof the study, and the significance of the study. In chapter two, a review of the previous related research is presented. The research methodology is presented in chapter three including the dataset, tools, and algorithms.In Chapter four, the prediction results obtained are presented and comparisons performed between the various algorithms. Also, the performance of the machine learning algorithms when recursive feature elimination with cross-validation and principal component analysis were applied to the dataset independently was examined. In Chapter five, conclusions were drawn from the results obtained. Furthermore, recommendations for further research were outlined.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Dimensionality Reduction

Dimensionality reduction is a data preprocessing operation which removes noisy, irrelevant, and redundant features to improve the performance metrics of a classifier and the time used to train and test the dataset (Velliangiri*et al.*, 2019). Dimensionality reduction can be performed in two ways. For the first way, the most relevant features in the original dataset are kept while the less relevant and redundant features are discarded. This technique is known as feature selection. For the second method, the redundancy of attributes in the dataset is exploited and finding new smaller attributes which are a combination of the original input attributes.This technique is known as feature extraction (Sorzano *et al.*, 2014.).

2.1.1 Feature selection

Feature selection refers to the process of determining the relevance of a feature to a problem aimed at removing redundant and irrelevant features in the dataset which in turn improves the performance and training time of the machine learning models(Mythily & Banu, 2017; Sorzano *et al.*, 2014.; Xie*et al.*, 2017). As a data preprocessing technique, feature selection has been used by researchers to improve the performance of machine learning models. As a result, the removal of redundant and irrelevant features enables the classifier to focus on the attributes of interests that enhance the classification of previously unseen data. Feature selection is also a type of dimension reduction that alleviates the dimensionality problem known as the curse-of-dimensionality(Xie *et al.*, 2017). The curse of dimensionality is a problem where error increases with an increase in the number of features. To overcome the curse-of-dimensionality problem, feature selection performs attribute subset generation and evaluation operations where a subset of the attributes from the dataset is formed and evaluated to either be optimal or not(Velliangiri *et al*., 2019). There are three types of feature selection techniques namely, filter, wrapper, and embedded(Mythily & Banu, 2017; Sorzano *et al*., 2014.; Velliangiri *et al*., 2019; Xie *et al*., 2017).

In filter feature selection, each feature is assigned a score and then ranked. Algorithms that are used to implement filter feature selection are information gain, Pearson's correlation, chi-square, and correlation coefficient. This method is fast and highly scalable with high dimensional data but it reduces a classifier's performance because it evaluates each dataset attribute independently(Mythily & Banu, 2017; Velliangiri *et al*., 2019).

The wrapper feature selection method forms a subset of features and trains the classifier on the feature subset. It uses objective functions which defines the metric of interest to determine the suitability of a subset of features.Based on the value of the objective function obtained, certain features are removed or added to the feature subset. Generally, wrapper feature selection is computationally expensive compared to filter methods, however, the wrapper method of feature selection guarantees the best feature subsets compared to filter methods (Mythily & Banu, 2017; Velliangiri *et al*., 2019). Algorithms that are based on the wrapper feature selection process includes backward feature elimination, recursive feature elimination, forward feature selection, genetic algorithm, and particle swarm optimization.

The embedded feature selection method combines the filter and wrapper methods which are in-built with the classifier. Decision trees is popular embedded feature selection method that serves as both a classifier and a feature selection model.

2.1.1.1 Recursive feature elimination with cross-validation (RFECV)

Recursive Feature Elimination with Cross-Validation (RFECV) is a wrapper feature selection algorithm that trains a classifier on all the features in a dataset before selecting subsets of the dataset's features at each iteration. RFECV then cross-validates the features chosen against the classifier to mitigate the stochastic nature of machine learning classifiers. RFECV was used to select features by recursively exploring smaller sets of features continuously in a cross-validation loop to obtain the optimal feature count(Velliangiri *et al.*, 2019).

2.1.2 Feature extraction

Feature extraction involves the creation of a new attribute set from the original dataset by decreasing the number of attributes to be processed while ensuring that relevant attributes in the dataset are captured(Velliangiri *et al.*, 2019). It constructs informative and non-redundant attributes from the original dataset by transforming data from a high dimensional space to a low dimensional space(Xie *et al.*, 2017).Various feature extraction techniques that exist include latent semantic analysis, principal component analysis (PCA), and independent component analysis (Velliangiri *et al.*, 2019).

2.1.2.1 Principal component analysis

Principal Component Analysis (PCA) is the most popular algorithm for dimension reduction which transforms a large set of variables into a smaller one but still contain most of the information in the previous set. Although the dimension reduction will reduce the accuracy or any other performance metric of interest of the classifier, the rationale behind this process is to sacrifice a little amount of accuracy or any other performance metric for simplicity. To obtain the principal components in the dataset, the dataset is first standardized(Velliangiri *et al.*, 2019).

Standardization refers to the scaling of values so that they fall within the same range. Since PCA is sensitive to the variances in variables which could lead to a biased result, variables must be scaled. The z-score normalization technique is applied to scale the variables in PCA accordingly. To perform z-score normalization, the value of the variable is subtracted from the mean value of the set of variables which is then divided by the standard deviation of the set of variables. Once the variables have been normalised, covariance matrix computation is performed.

The covariance matrix computation process is aimed at understanding how the variable of the dataset varies from the mean value. That is, this process seeks to establish the relationship between the normalized variables. With a covariance matrix, highly correlated variables are revealed. A variable is highly correlated with another if it has as much information as the other variable. This means that both variables contain redundant information. The covariance matric is a symmetric matrix of the dataset dimension. For instance, if the number of attributes in the dataset is p, then, the covariance matrix is represented by a $p \times p$matrix. For the covariance result, a positive covariance value indicates that both variables increase or decrease together. Conversely, a negative covariance value indicates that as one variable increases, the other decreases. Next, the eigenvectors and eigenvalues of the covariance matrix are computed to identify the principal components(Jaadi, 2021).

Eigenvector of a linear transformation is a nonzero vector that changes at most by a scalar factor when that linear transformation is applied to it.In the covariance matrix, the eigenvectors are the directions of the axes where there is the most variance. The eigenvalue is the factor by which the eigenvector is scaled.That is, the eigenvalue is the coefficients attached to eigenvectors that give the amount of variance carried in each principal component. The principal components are obtained in their order of significance by ranking the eigenvectors from highest to lowest in order of the corresponding eigenvalue. Principal components are new variables formed from a linear combination of the initial variables(Jaadi, 2021).

According to Sorzano *et al*. (2014) PCA can be represented mathematically as thus, given a set of observations x, with dimension M which lies in ℝ, PCA is the standard method for obtaining the best subspace of a dimension m. PCA is based on orthogonal direction search which explains the variance of the data. The dimensionality reduction problem can be expressed as a problem of obtaining the m orthonormal directions $w_i$which minimizes the representation error given in Equation 2.1 according toSorzano *et al*. (2014):

$$J_{PCA} = E\{\|x - \textstyle\sum_{i=1}^{m} \langle w_i - x \rangle w_i \|^2\} \tag{2.1}$$

where x is a set of observations and w_i represents the orthonormal directions. In the objective function expressed in Equation 2.1, the reduced vectors,χ, are the projections described in Equation 2.2 as:

$$\chi = (\langle w_1, x \rangle, \dots, \langle w_m, x \rangle)^t \tag{2.2}$$

Equation 2.2 can be reduced further to Equation 2.3

$$\chi = W^t x \tag{2.3}$$

where W is a $M \times m$ matrix whose columns are the orthonormal directions w_i or equivalently $W^t W = I$. The approximation of the original vectors is given in Equation 2.4 as:

$$\hat{x} = \textstyle\sum_{i=1}^{m} \langle w_i - x \rangle w_i \tag{2.4}$$

which is equivalent to $\hat{x} = W\chi$. When the input vectors are standardized, the objective function solution J_{PCA} is given by m eigenvectors associated with the largest $m$eigenvalues of the covariance matrix of x given in Equation 2.5 as:

$$C_x = \frac{1}{N} XX^t \tag{2.5}$$

where C_x is a $M \times M$ matrix of M eigenvalues. If the eigenvalue decomposition of the input covariance matrix is $C_x = W_M \Lambda_M W_M^t$, then the feature vectors are constructed as given in Equation 2.6.

$$\chi = \Lambda_m^{-\frac{1}{2}} W_m^t x \tag{2.6}$$

where Λ_m is a diagonal matrix with the m largest eigenvalues of the matrix Λ_M and W_m are the corresponding m columns from the eigenvectors matrix W_M. All features can be constructed by projecting the whole matrix$X$. The ith feature is the projection of the input vector x onto the ith eigenvector which is given in Equation 2.7 as:

$$\chi_i = \lambda_i^{-\frac{1}{2}} w_i^t x \tag{2.7}$$

2.2 Machine Learning Algorithms for Predicting Loan Defaults

Machine learning has been applied to various fields including agriculture. According to Liakos*et al.*(2018), machine learning has been applied to crop, livestock, water and soil management through the use of data obtained from sensors. Consequently, recommendations have been drawn from the results obtained through machine learning which has improved the yields of crop and sustained the health of livestock. Machine learning consists of groups of algorithms based on the learning type and another group based on the learning model. For the learning type, it could either be a supervised, unsupervised, or semi-supervised learning model. For the learning model, it could either be classification, regression, or clustering based on the problem being solved(Liakos *et al.*, 2018). In supervised learning, a training dataset is provided for the machine learning algorithm to learn from and establish patterns that will enable it to classify previously unseen data effectively. Supervised learning is synonymous with classification. Unlike supervised learning, unsupervised learning identifies relationships between the instances of the dataset and uses the relationships identified to classify previously unseen data.Unsupervised learning is synonymous to clustering. In semi-supervised learning, training of the machine learner is performed using labelled and unlabelled datasets (Agarwal, 2014). In this work, prediction of agricultural loan defaulters is a classification problem therefore, classifiers are used. Several machine learning classifiers exist which includes decision trees, support vector machines, random forest, gradient boosting, adaptive boosting, and logistic regression. These classifiers are the most utilized classifiers in the domain of credit risk management.

2.2.1 Decision trees

A decision tree is a tree structure where each node represents a test on the value of an attribute, each branch denotes the outcome of the test, and tree leaves represent classes or class distributions. In decision trees, the dataset is progressively organized into smaller homogenous subsets while generating the associated tree graph (Agarwal, 2014; Velliangiri *etal*., 2019).

2.2.2 Support vector machine

In support vector machine classification, the training data is projected into a higher dimension where the hyperplane that separates the data by classes using essential training tuples called support vectors. It uses nonlinear mapping to transform the training data into a higher dimension. Using global optimization, the support vector machine deals with overfitting problems which makes them suitable for classification, regression, and clustering tasks (Agarwal, 2014; Velliangiri *et al*., 2019).

2.2.3 Random forest

Random forest is an ensemble classifier that consists of other individual decision trees which form a forest. The individual decision trees are generated using a random selection of attributes at each node to determine the split. Random forests are robust to errors and outliers. Furthermore, the error of generalization for a forest converges once there is a large number of trees in the forest(Agarwal, 2014). It fits several decision tree classifiers on different subsets of the dataset and uses averaging to mitigate over-fitting and improve the accuracy

2.2.4 Gradient boosting

Gradient boosting is an ensemble machine learning technique that mostly uses decision trees which are weak learners. In gradient boosting, the classifier consecutively fits new models to provide a more accurate estimate of the response variable (Natekin & Knoll, 2013). Gradient boosting builds an additive model in a forward stage-wise fashion.

2.2.5 Adaptive boosting (AdaBoost)

Adaptive Boosting (AdaBoost) is a popular boosting algorithm where the output of other weak learning algorithms is combined into a weighted sum that represents the output of the final output of the boosted classifier. Here, for the training dataset, if an instance in the dataset is misclassified, the weight is increased but if an instance in the dataset is correctly classified, the weight is reduced(Agarwal, 2014).Adaboost begins by fitting the classifier on the original dataset. Subsequently, additional copies of the classifier are fitted on the same dataset and weights of incorrectly classified instances are adjusted.

2.2.6 Logistic regression

Logistics regression is a statistical model which utilizes logistic function to model a binary dependent variable(Agarwal, 2014). It measures the relationship between the categorical dependent variable and other variables by estimating probabilities using a logistic function. The binary logistic regression model was used in this work. The liblinear solver for the optimization problem was employed because liblinear works well for small datasets and it handles the one-versus-rest schemes.

2.3 Review of Related Works

The need to harness opportunities using resources that are not available at the moment usually informs the decision to seek loans, either from individuals or financial institutions. Based on the need and required amount, a choice is made between obtaining loans from individuals within one's social circle or established financial institutions. On the one hand, receiving loans from individuals confers benefits such as low to no interest rate on the borrowed cash. On the other hand, however, the amount borrowed may be small relative to the needed amount. Therefore, individuals approach financial institutions for credit. Compared to the peer-to-peer lending option which is based on a social trust model, financial institutions endeavour to limit the risk of the loan through various techniques including a request for collateral with a value greater or equal to the loan amount. Among other techniques is the use of machine learning algorithms to predict loan defaulters from previous loan datasets to minimize risk. Researchers have examined the prediction of commercial bank loan defaulters using various statistical and machine learning methods, therefore, this research focuses on applying the aforementioned machine learning techniques in predicting agricultural loan defaulters.

A credit score model for airtime loans using machine learning was postulated by Dushimimana*et al*. (2020) using a dataset obtained from ComzAfrica. In the research, machine learning algorithms – logistic regression, decision tree, and random forest – were applied to the ComzAfrica dataset of 1 January 2016 to 30 June 2017. The in-sample analysis of the algorithms yielded a uniform accuracy of 99.1% and specificity of 0.2%, 0.0%, and 0.8% for logistic regression, decision tree, and random forest algorithms respectively. Although the accuracy of the algorithms was high, the low specificity obtained indicates that the classifier incorrectly predicts default when considering those that defaulted.

Datkhile *et al*. (2020) utilized kaggle credit dataset with 12 attributes to predict loan default using logistic regression, naïve bayes, decision trees, and random forest algorithms. The results showed that logistic regression had the highest accuracy of 93.777%. The accuracy of random forest, naïve bayes, and decision trees were 93.44%, 89.86%, and 89.51%. Apart from the accuracy of the classifiers, other metrics were not measured.

Similarly, a dataset with 64,000 tuples and 14 attributes was used in forecasting loan default in the research by Patel*et al*. (2020). The algorithms: logistic regression, gradient boosting, catboost classifier, and random forest, achieved accuracy and precision of 14.96% and 49%, 84.04% and 85%, 84.05% and 85%, and 83.51% and 86% respectively. All the algorithms except logistics regression achieved F1-score of 91%.

In line with the work by Patel *et al*. (2020), logistic regression was used to determine the likelihood of loan default in Bangladesh (Aslam*et al*., 2020). The authors were able to establish that logistics regression classifies 93.30% of the cases. Similarly, Elakkiya*et* *al*. (2020) used logistic regression model to predict loan defaults. The authors were able to achieve accuracy, precision, recall, and f1−score of 88.83%, 91.07%, 58.47%, and 71.22% respectively.

Coşer*et* al. (2019) developed predictive models to assess loan risk using LightGBM, XGBoost, Logistic Regression and Random Forest. Random forest obtained the best results with an Area Under Receiver Operating curve of 89%. The use of logit model to predict micro-loan default in the LendingClub dataset was examined by Deng (2019). The feature selection process which was applied to the dataset selected 20 features with the greatest impact using correlation coefficient analysis. The logistics regression model achieved an accuracy of 92.9%. Other metrics were not evaluated.

Bayraci and Susuz (2019) applied deep neural network to predict loan default which was compared with logistic regression, decision tree, naïve Bayes, and support vector machine algorithms. The authors used two distinct datasets with 79254 instances obtained from a medium-sized Turkish bank. For the loan performance data, the proposed deep neural network model achieved a weighted accuracy of 77.98%. Logistic regression, decision tree, naïve Bayes, and support vector machine algorithms achieved weighted accuracy of 77.31%, 70.05%, 78.14%, and 57.04% respectively. The percentage of misclassified good loans and bad loans performance for deep neural network are 10.20% and 25.95% respectively. For the loan application data to discriminate between the creditworthy and non-creditworthy applicants, the proposed deep neural network model achieved a weighted accuracy of 85.69%. Logistic regression, decision tree, naïve Bayes, and support vector machine algorithms achieved weighted accuracy of 78.01%, 82.34%, 77.93%, and 75.25% respectively. The percentage of misclassified good loans and bad loans applications for deep neural network is 15.45% and 13.92% respectively.

In a bid to improve the performance of loan default prediction methods, Al-Qerem*et al.* (2019) explored comprehensive pre-processing, extraction, and selection of features in the dataset. The enhancement approach which utilized information gain, genetic algorithm and particle swarm optimization for feature selection was tested using naïve Bayes, decision tree, and random forest classifiers. It was established that the data pre-processing methods improved classification accuracy and model performance.

OptiML was used by Zoran (2019) to forecast credit non-payments using the dataset from a microcredit organisation. Three models were shortlisted for evaluation after executing OptiML on the dataset: decision forest, neural network, and a logistic regression model. The decision tree achieved accuracy, precision, and F-measure of 94.6%, 69.5%, and 0.0596 respectively. Similarly, the neural network model achieved accuracy, precision, and F-measure of 82.1%, 15.5%, and 0.2396 respectively. Furthermore, the logistic regression model achieved accuracy, precision, and F-measure of 94.7%, 66.0%, and 0.0463 respectively.

Fu *et a*l.(2020) focused on forecasting loan defaults in online lending peer-to-peer systems using bidirectional long short term memory (BiLSTM). The dataset comprised over 440000 online comments on about 6000 online P2P lending companiesfrom Wangdaizhijia in China. The performance of the proposed model was compared with support vector machine, decision tree, deep neural network, and text convolutional neural network. The proposed method achieved precision, recall and f1 scores of 0.7964, 0.7740, and 0.8034 respectively.

Madane &Siddharth (2019) applied logistic regression, random forest, decision tree, adaboost, XGboost, artificial neural network and support vector machine algorithms to predict loan defaults. Also, the Synthetic Minority Oversampling Technique (SMOTE) was employed to treat the imbalance between classes for the response variable. It was observed that XGBoost without implementation of SMOTE obtained the best result.

Kim and Cho (2019) combined label propagation and transductive support vector machine (TSVM) with Dempster–Shafer theory for accurate default prediction of social lending using unlabelled data. The experiment was performed using the Lending Club dataset. The proposed method achieved accuracy and f1-score of 76.79% and 86.47% respectively. In another perspective on loan repayment, ascertaining the likelihood of repayment of a credit card loan was examined by Ma (2020). The author applied XGBoost model to a dataset with 30,000 samples of credit-card billing information and repayment information. The proposed model achieved an Area Under Receiver Operating Curve (AUC) of 0.779.

Semiu and Gilal (2019) applied a boosted decision tree model for forecasting loan default in peer-to-peer lending communities using the publicly available United States small business administration dataset and the Imperial College London Kaggle competition dataset. The dataset which consists of 899,164 data instances was used in the 80:20 ratio for training and testing. On applying decision tree and boosted decision tree model to the dataset, 99% and 98% accuracy were recorded.

Similarly, Zhou*et al.* (2019) predicted loan default in a peer-to-peer lending platform using a heterogeneous ensemble decision tree model based on gradient boosting decision trees, extreme gradient boosting, and light gradient boosting machine. The ensemble method obtained sensitivity, specificity, f1-score, and accuracy values of 0.9596, 0.1589, 0.8615, and 0.7185.

A Taiwan credit dataset was employed in forecasting loan defaults by Motwani*et* al. (2018). The classification task was performed using the bagging ensemble method with

REP tree algorithm, linear regression, and decision stump. The proposed work obtained an accuracy of 81% when REP Tree was used compared to the base learners. In another work, Motwani*et al.*(2018) examined the calculation of a bank's customer credit worthiness using Microsoft Azure machine learning studio. The proposed method was compared against three algorithms: Bayes point, logistic regression, and decision tree. The proposed method achieved accuracy, true positive, recall, and prediction rate of 82.20%, 1360instancesss, 0.411, and 0.110 respectively.

Nalić and Švraka (2018) presented a credit scoring model used by two microfinance institutions: one in Bosnia, the other Herzegovina. Data preprocessing was performed using Oracle data miner on the dataset which has 87531 records with over 60 attributes. The Generalized Linear Models (GLM) algorithm in the Oracle data miner software was used to perform the classification task. Results obtained showed that GLM achieved an overall accuracy of 98.2046% and average accuracy of 98.7185%. The review of related work is summarised in table 2.1.

Table 2.1: Review of Related Work

S/N	Author/Year	Technique	Strengths	Weaknesses
1	Dushimimana *et al*. (2020)	Logistic regression, decision tree, and random ssforest	99.1% accuracy was recorded for all the algorithms used	Low specificity indicates that the classifier incorrectly predicts default when considering those that defaulted
2	Datkhile *et al*. (2020)	Logistic regression, naïve Bayes, decision trees, and random forest algorithms	Recorded high accuracy for random forest, naïve Bayes, and decision trees of93.44%, 89.86%, and 89.51% respectively	Evaluated accuracy only. Did not consider other performance metrics
3	Patel *et al*. (2020)	Logistic regression, gradient boosting, catboost classifier, and random forest	Logistic regression achieved an F1-score of 91%	Logistic regression achieved low accuracy and precision of 14.96% and 49% respectively
4	Aslam *et al*. (2020)	Logistic regression	Logistics regression classifies 93.30% of the cases	Did not evaluate other performance metrics
5	Elakkiya *et al*. (2020)	Logistic regression	Achieved accuracy, precision, recall, and f1−score of 88.83%, 91.07%, and 71.22% respectively	A low recall score of 58.47% was recorded which means that loan defaulters were not identified adequately
6	Coşer *et al*. (2019)	LightGBM, XGBoost, Logistic Regression and Random Forest	Achieved Area under the Receiver Operating Curve of 89%	A low recall score of 67.6% was recorded

Table 2.1: Review of Related Work (continued)

7	Bayraci and Susuz (2019)	Deep neural network, logistic regression, decision tree, naïve Bayes, and support vector machine algorithms	Achieved good accuracy rates for deep neural network, logistic regression, decision tree, and naïve Bayes	Misclassification rate ranging from 10% to 25% for good and bad loans was obtained
8	Al-Qerem *et al*. (2019)	Naïve Bayes, decision tree, and random forest classifiers	It was established that feature selection improves the performance metrics of loan default classifiers	On average, feature selection improved performance metrics by 3%
9	Zoran(2019)	Decision forest, neural network, and a logistic regression model	Accuracy of 94.6%, 82.1%, 94.7% for the decision tree, neural network, and logistic regression respectively was recorded	Low F1-score was recorded which highlights the low recall values
10	Fu et al. (2020)	Bidirectional long short term memory, support vector machine, decision tree, deep neural network, and text convolutional neural network	Achieved precision, recall and f1 scores of 0.7964, 0.7740, and 0.8034 respectively	Focused on using comments which may not be adequate when comments are few
11	Madane and Siddharth(2019)	Logistic regression, random forest, decision tree, adaboost, XGboost, artificial neural network and support vector machine	XGBoost without implementation of SMOTE obtained the best result	XGBoost is likely to overfit the training data
12	Kim and Cho(2019)	Label propagation and transductive support vector machine (TSVM) with Dempster–Shafer theory	Achieved an accuracy and f1-score of 76.79% and 86.47% respectively	Only accuracy and f1-score were the performance metrics highlighted

Table 2.1: Review of Related Work (continued)

13	Ma(2020)	XGBoost	Achieved an Area Under Receiver Operating Curve (AUC) of 0.779	Other performance metrics were not evaluated
14	Semiu & Gilal (2019)	Boosted decision tree	Achieved accuracy of 99% and 98% fordecision tree and boosted decision tree model respectively	Apart from accuracy, other performance metrics were not investigated
15	Setiawan *et al*. (2019)	Extremely randomized tree and random forest	Extremely randomized tree achieved an accuracy of 64% and better execution time up to 46% compared to random forest	The accuracy values obtained are not competitive when compared with the findings of other researchers
16	Zhou *et al*. (2019)	Used a heterogeneous ensemble decision tree model based on gradient boosting decision trees, extreme gradient boosting, and light gradient boosting machine	The ensemble method obtained sensitivity, f1-score, and accuracy values of 0.9596, 0.8615, and 0.7185	The low specificity of 0.1589 means that most loan non-defaulters are denied access to loans
17	Motwani *et al*. (2018)	Bagging ensemble method with REP tree algorithm, linear regression, and decision stump	The proposed method achieved accuracy, true positive, recall, and prediction rate of 82.20%, 1360, 0.411, and 0.110 respectively.	Low recall and prediction rate of 0.411 and 0.110 respectively
18	Nalić and Švraka (2018)	Oracle data miner	Achieved an overall accuracy of 98.2046%	Possibility of overfitting

CHAPTER THREE

3.0 RESEARCH METHODOLOGY

3.1 Proposed Approach

The dataset undergoes the z-score normalization before feature selection and feature extraction were performed on the dataset. The selected and extracted features were classified using random forest, support vector machine, gradient boosting, ada boosting, and logistic regression. The choice of these classifiers is premised on the use of the classifiers by Dushimimana *et al.*(2020) and Patel *et al.*(2020). The approach used in this research is illustrated in figure 3.1.

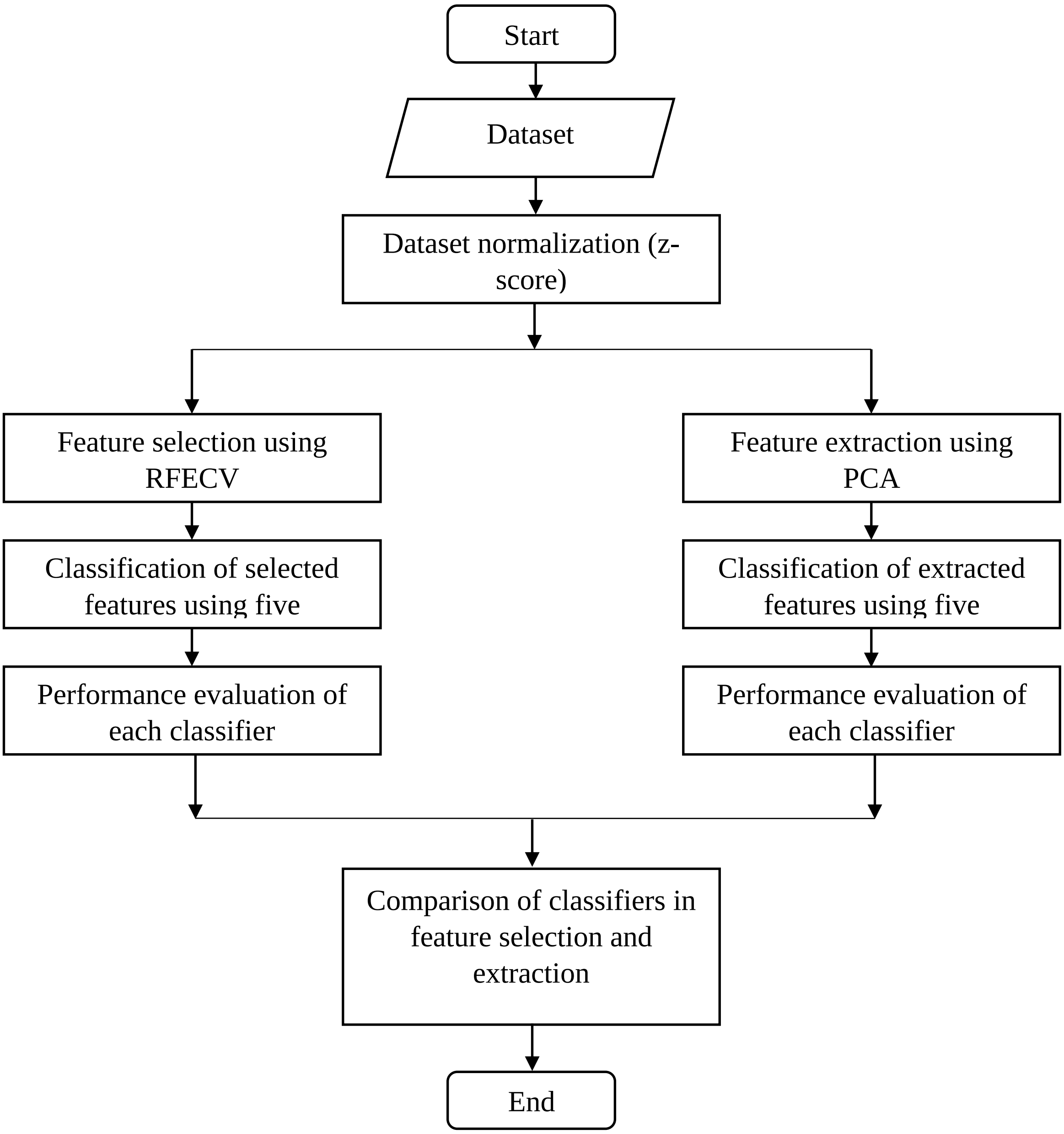
Start
Dataset
Dataset normalization (z-score)
Feature selection using RFECV
Feature extraction using PCA
Classification of selected features using five
Classification of extracted features using five
Performance evaluation of each classifier
Performance evaluation of each classifier
Comparison of classifiers in feature selection and extraction
End

Figure 3.1: Proposed Approach

3.2 Dataset

The dataset used in this work was obtained from a financial institution that liaised with farmers in Lavun local government area of Niger state in Nigeria. The financial institution facilitated the issuance of agricultural loans to farmers in Lavun who predominantly cultivate rice. The dataset contains 174 unique loan instances with labels indicating a loan default. Due to the number of attributes, 60 features, captured in the dataset relative to the number of unique instances, the dataset was subjected to feature selection and feature extraction process to enhance the performance metrics. The dataset adhered to a uniform distribution pattern. That is, there wasan equal number of loan default and non-defaulting instances in the dataset.Furthermore, in performing the classification task, the dataset was split into training, testing, and validation datasets in the ratio of 60:20:20. That is, 60% of the dataset was used for training, 20% for testing, and 20% for validation. Before feature selection and extraction were performed, data normalization was performed.

3.3 Data Normalization

Data normalization refers to the process of giving the same weights to variables in an observation. That is, the variables in the observation or dataset are transformed into smaller ranges to prevent dominance of other variables by a larger value. Normalization is crucial in classification to prevent attributes with initially large ranges from overshadowing attributes with initially small ranges. Methods for data normalization include min-max normalisation, z-score normalisation, and normalization by decimal scaling. However, since the feature extractor, PCA, and one of the classifiers, support vector machine, both require data normalization, z-score normalisation was performed on the dataset. Z-score normalisation has been proven to work well with PCA and support vector machine.

3.3.1 Z-Score normalization

Z-score normalisation standardizes the values of an attribute by subtracting each attribute from the mean value of that attribute and dividing the result by the standard deviation of that attribute. Given a value v, the value$v^{\prime}$ to be obtained on performing z-score normalisation can be expressed in Equation 3.1 as:

$$v^{\prime} = \frac{v-\mu}{\sigma} \qquad (3.1)$$

where μ is the mean value of the attribute and σ is the standard deviation.

3.4 Feature Selection using RFECV

Given a large number of attributes in the dataset, it is imperative to perform dimensionality reduction to improve the performance of the machine learning classifiers.According to Nalić and Švraka(2018), performing feature selection dimensionality reduction technique generated reliable credit scoring models in their research. Since dimensionality reduction involves either feature selection or extraction, both were selected for use in this research to ascertain which method works best on the agricultural loan dataset due to the large number of attributes in the dataset. While feature selection is focused on picking a subset of features in the dataset, feature extraction compresses the data as it preserves the most relevant information. For feature selection, Recursive Feature Elimination with Cross-validation (RFECV) was used. RFECV trains a classifier on all the features in a dataset before selecting subsets of the dataset's features at each iteration. Then cross-validation of the chosen features is performed to ensure that consistent output is obtained. RFECV recursively explores smaller sets of features until it determines the set of features that best gives good performance according to the metric. In this research, random forest was used as the estimator for the RFECV. The choice of random forest, as an estimator, was based on its ability to perform classification tasks on datasets without any need for data normalization. For cross-validating, the dataset, 5-fold cross-validation was performed. Algorithm 3.1 describes the operations of RFECV.

Algorithm 3.1: Recursive Feature Elimination with Cross-Validation

```
Train random forest classifier on the training set of data
Perform 5-fold cross-validation
Calculate variable rankings
For each subset size $S_i$, I = 1,2,…,S do
    Keep $S_i$ most important variables
    Calculate random forest classifier performance
    Perform 5-fold cross-validation
End
Calculate the profile performance over $S_i$
Determine the appropriate numbers of features
```

3.5 Feature Extraction using Principal Component Analysis

Principal Component Analysis (PCA) was applied using sklearn in python. The PCA uses linear dimensionality reduction using singular value decomposition of the data to project it to a lower-dimensional space. The principal components selected explained 95% of the variance. The principal components were then saved to an excel file for classification tasks to execute. Therefore, 31 principal components that explained 95% of the variance were used.

3.6 Classification

The classification of the selected features and extracted features were performed independently. Random forest, support vector machine, gradient boosting, adaptive boosting, and logistic regression are the classifiers used to predict loan defaulters in the dataset obtained. Since random forest, gradient boosting, adaptive boosting, and logistic regression are not sensitive to unscaled values, the training data was not scaled(Pedregosa *et al*., 2011). However, the training data for support vector machine was scaled because the classifier is sensitive to varying large values. The radial basis function kernel support vector machine was used to detect loan defaulters better. The C parameter for the radial basis function kernel support vector machine was set to 3 while the gamma parameter was set to 0.01. For the logistic regression, the liblinearsolver was selected because it works well with small datasets and handles one-versus-rest schemes properly(Pedregosa *et al*., 2011). That is, it handles the detection of one class of interest in a dataset against all others.For the random forest classifier, the number of decision trees (estimators) was set to 100 with a maximum depth of 5. For the gradient boosting ensemble classifier, the number of decision trees was also set to 100 while the maximum tree depth was set to 6. For the adaptive boosting classifier, the number of estimators was set to 50.

Furthermore, 10-fold cross-validation was applied to each classifier to measure how well each model will generalize and to identify possible model overfitting. Cross-validation is a model validation technique that ensures that machine learning models can predict unseen data. The 10-fold cross-validation technique was selected because it has been established by literature that it results in less biased models.

3.7 Performance Metrics

The performance of the classifiers used in this research defines how well agricultural loan defaulters in Lavun local government area are identified from non-defaulters. The performance metrics used to measure the suitability of each classifier are, accuracy, Area Under the Receiver Operating Characteristic Curve (AUC), precision, recall, and f1-score.

3.7.1 Accuracy

The accuracy of a classifier refers to the number of correctly classified instance in the dataset. That is, the sum of true positive instances with that of the true negative instances divided by the total number of instance. Here, true positive (TP) refers to the number of loan default that was correctly classified as loan defaults while a true negative (TN), refers to the number of non-defaulting loans that were correctly classified as non-default loans. The formula for calculating accuracy is given in Equation 3.2.

$$Accuracy = \frac{TP+TN}{Number\ of\ instances} \tag{3.2}$$

3.7.2 Area under the receiver operating characteristic curve

Area Under the Receiver Operating Characteristics (AUC) refers to the degree of the classifier’s separability based on the Receiver Operating Characteristic (ROC) curve. ROC curves, a probabilistic curve, are visual tools used in comparing classification models. It shows the trade-off between true positive rate and false-positive rate. Here, the true positive rate refers to the rate at which a defaulting loan is classified as such while the false positive rate defines the rate at which non-defaulting loans are classified as loan defaults. The AUC values ranges from 0 to 1. Higher AUC values within the 0-1 range means the classifier is better at predicting loan defaults and non-defaulting loans appropriately.

3.7.3 Precision

Precision, a measure of exactness, refers to the percentage of correct predictions among the test data. It measures the exactness of the classifier. The formula used in calculating precision is given in Equation 3.3. Here, false positive (FP) refers to the non-defaulting loans mistakenly classified as defaulting.

$$Precision = \frac{TP}{TP+FP} \tag{3.3}$$

3.7.4 Recall

Recall, also known as sensitivity, is defined as the number of positive cases that were correctly identified. It measures the completeness of the classifier. The formula for calculating recall is given in Equation 3.4. Here, false negative (FN) refers to the defaulting loans which were mistakenly classified as non-defaulting.

$$Recall = \frac{TP}{TP+FN} \tag{3.4}$$

3.7.5 F1-Score

F1-score, also known as f-score, is the harmonic mean of the precision and recall score. In other words, it conveys the balance between the precision and the recall of a classifier. A model with the best performance shows a maximum f1-score. The formula for calculating the f1-score of a model is given in Equation 3.5.

$$\frac{2 \text{ x precision x recall}}{\text{precision+recall}} \tag{3.5}$$

CHAPTER FOUR

4.0 RESULTS AND DISCUSSIONS

4.1 RFECV Feature Selection Result

For the RFECV feature selection, random forest was used as the estimator and the accuracy of the random forest classifier was the objective function of the feature selection process. The RFECV feature selection process selected 44 features as the features of interest out of the 60 features in the dataset. All the features of the dataset are presented in Table 4.1 and the selected features are denoted by an asterisk (*).

4.2 Classification of RFECV Selected Features

The machine learning algorithms selected for the classification process were applied to the dataset which had the selected features. The performance of the models was measured using accuracy, precision, recall, AUC, and f1-score metrics. Also, the receiver operating characteristic curve was plotted to show the performance of each model at all classification thresholds. The confusion matrix for each classifier is also presented in figure 4.7, 4.8, 4.9, 4.10, and 4.11. Table 4.2 shows the values obtained by each classifier for each performance metric while figure 4.1 illustrates the values represented in Table 4.2.

Table 4.1: All Features of the Dataset with the Selected Features

Sex	Yield sale to offtakers*	Yield sale to farmgate*	Yield sale to self*
Marital status*	Yield wasted*	Production method*	Farming cost*
Age range*	Awareness of credit*	Challenge accessing credit*	Count of credit access to farm
Farming regularity	Challenge repaying credit*	Bank account	Type of account*
Farming duration*	Account opening facilitator*	Account opening date*	Reason for opening account*
Farming system	Frequency of saving*	Withdrawal frequency*	Number of dependants*
Association link*	Household size*	Male children between 6 to 18 years*	Female children between 6 to 18 years*
Education level	Number of male children attending school*	Number of female children attending school*	Hunger due to inadequate food*
Farm size*	Main source of income	Total income per month*	Number of full time female workers*
Cultivation interval	Number of part time male workers*	Number of part time female workers*	Number of full time male workers
Crop type	Type of farm animals*	Food shortage*	Drinking water treatment*
Locality seasons*	House roofing material*	House building material*	Access to electricity*
Reason for not cultivating multiple seasons*	Access to agricultural insurance	Awareness of agricultural insurance	Account opening balance at start of season*
Season 1 yield gain	Access to health facilities	Toilet facility*	Mechanized farm tools*
Season 2 yield gain	Personal possession	Source of farm funding*	Media accessed by farmers*

Table 4.2: Classifier Performance Evaluation for RFECV Selected Features

Classifiers	Accuracy	AUC	Precision	Recall	F1-score
Gradient Boosting	88.57%	0.875	86.36%	95%	90.48%
Ada Boost	80%	0.808	88.24%	75%	81.08%
Random Forest	80%	0.809	89.47%	77.27%	82.93%
SVM	80%	0.806	73.68%	87.5%	79.99%
Logistic Regression	82%	0.845	72.22%	92.86%	81.25%

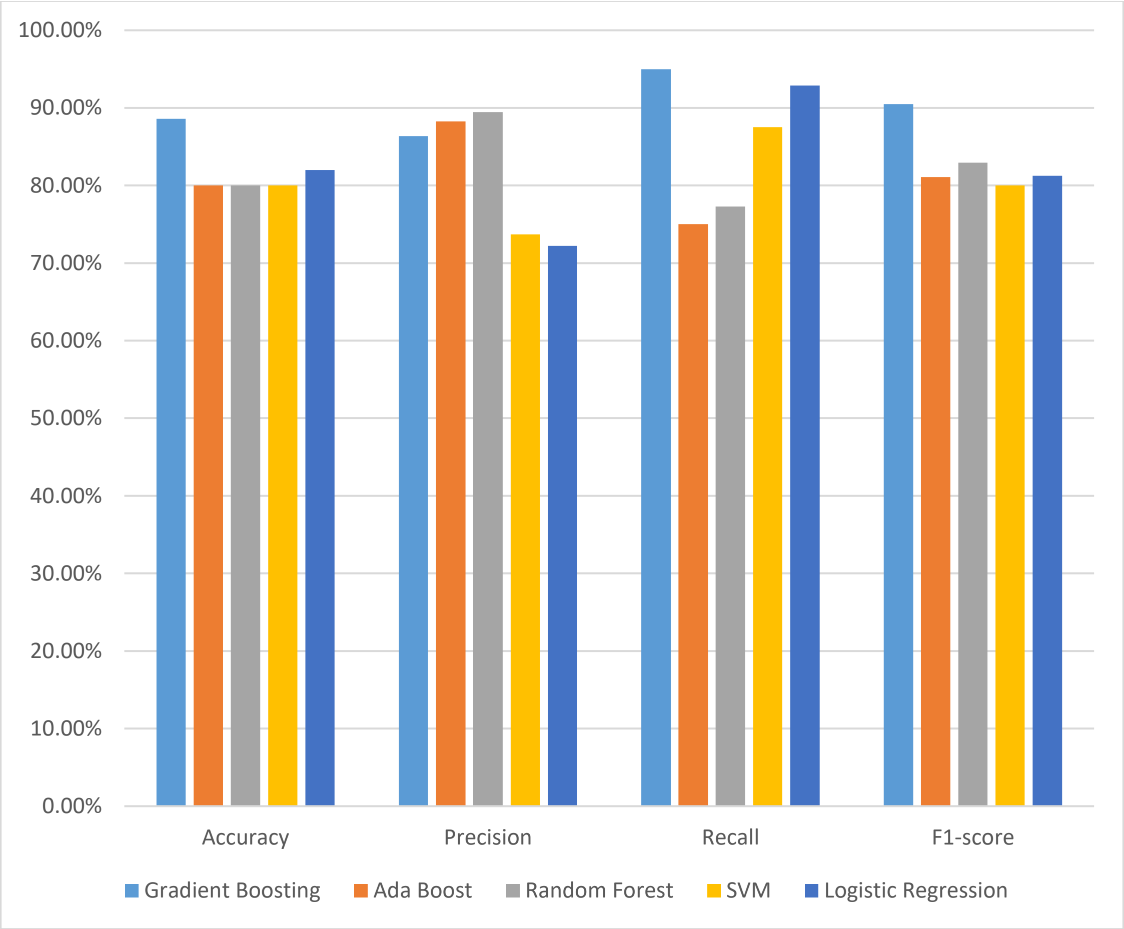

Figure 4.1: Performance of Classifiers on RFECV Selected Features

For the model accuracy, gradient boosting and logistic regression model obtained the highest accuracy of 88.57% and 82% respectively. The other models – Ada boost, random forest, and SVM – obtained an accuracy of 80%. Based on the accuracy of the models examined, gradient boosting was able to classify most of the test instances correctly compared to the other four models.

Similarly, gradient boosting and logistic regression model achieved the highest AUC scores of 0.875 and 0.845 respectively. This means that the gradient boosting and logistic regression model had better separability than ada boost, random forest, and SVM. In other words, the models were able to tag more defaulting loans as defaults and non-defaulting loans as non-default compared to the other models.

Interestingly, random forest and ada boost models attained the best precision scores of 89.47% and 88.24% respectively. That is, the random forest and ada boost models classified a lesser number of non-defaulting loans as defaults (false positive) while predicting a greater number of defaulting loans as defaults (true positive). Although logistic regression and gradient boosting models had the best accuracy and AUC scores, they performed poorly in labelling non-defaulting loans. In other words, logistic regression and gradient boosting models labelled more non-defaulting loans as defaults compared to random forest and ada boost models.

Gradient boosting and logistic regression models proved efficient in predicting agricultural loan defaults by attaining recall scores of 95% and 92.86% respectively. In other words, gradient boosting and logistic regression models were able to predict agricultural loan defaulters better than other classifiers. SVM followed with a recall score of 87.5%. Random forest and Ada boost trailed behind with recall scores of 77.27% and 75% respectively. When the recall scores and precision scores of random forest and Ada boost models are juxtaposed, it can be deduced that although the models classified a lesser number of non-defaulting loans as defaults, they could not identify agricultural loan defaulters effectively.

Consequently, the f1-score of 90.48% was obtained by gradient boosting model, thus, making the model the most effective in predicting agricultural loan defaulters. Random forest which performed averagely in most of the metrics, apart from the precision metric, proved to be the second most effective model with an f1-score of 82.93%. Logistic regression model positioned itself as a competitive model by obtaining an f1-score of 81.25%. Ada boost and SVM models attained f1-scores of 81.08% and 79.99% respectively. The ROC curves for ada boost, gradient boosting, logistic regression, random forest, and SVM are presented in figure 4.2, 4.3, 4.4, 4.5, and 4.6 respectively.s

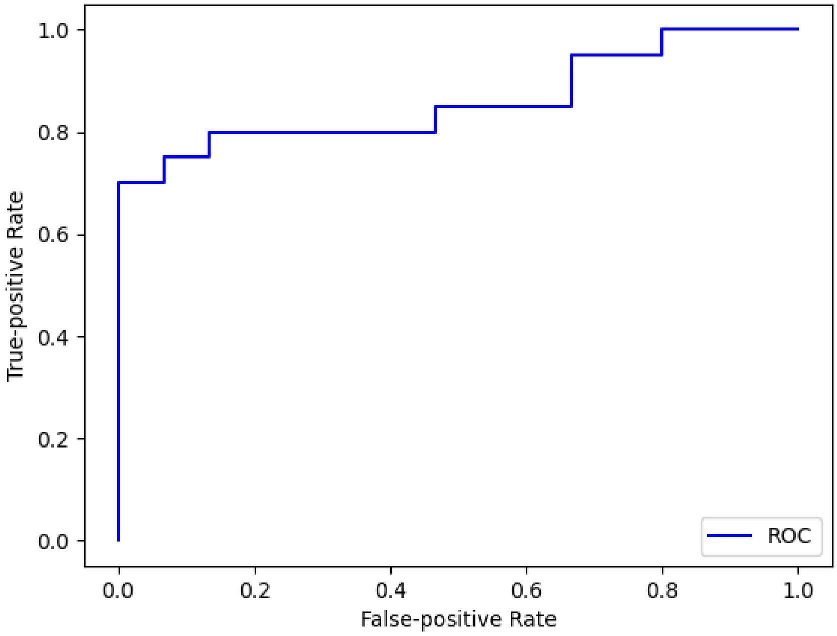

Figure 4.2: Adaptive Boosting RFECV ROC Curve

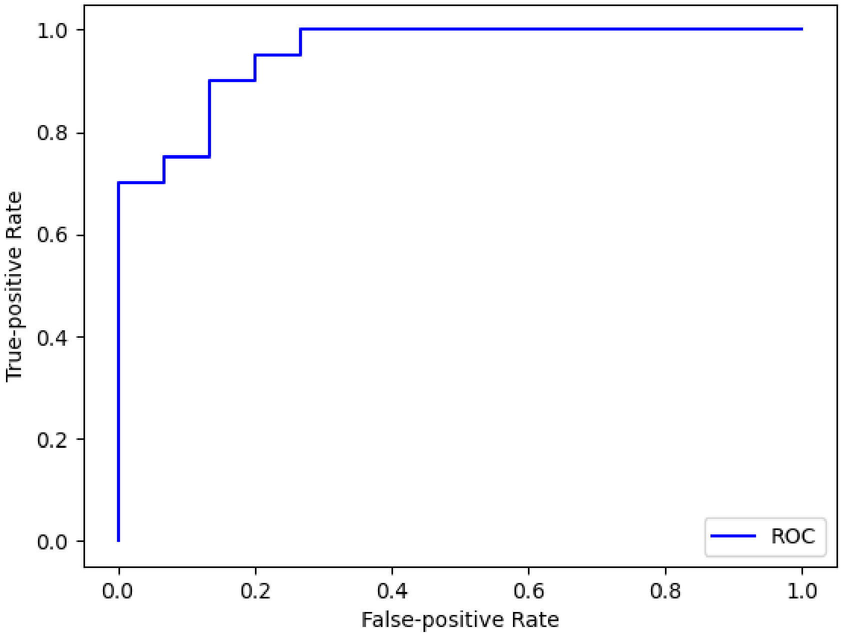

Figure 4.3: Gradient Boosting RFECV ROC Curve

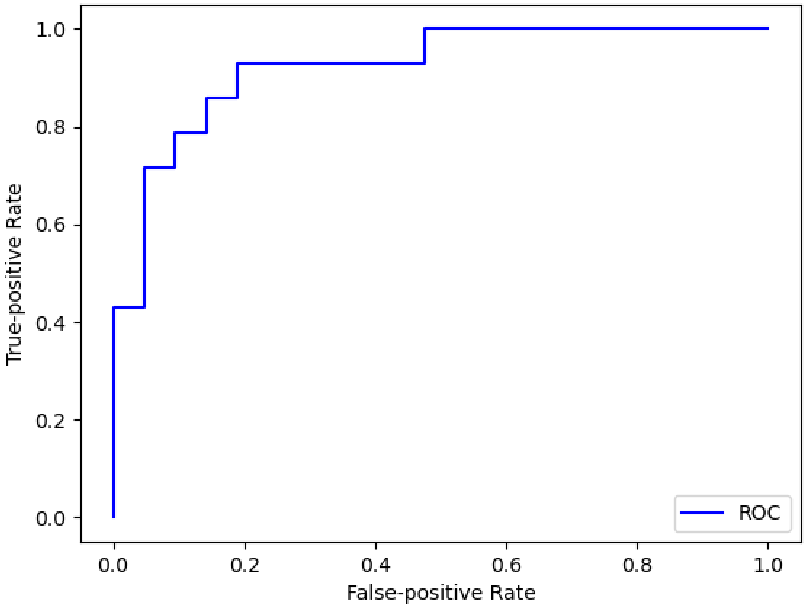

Figure 4.4: Logistic Regression RFECV ROC Curve

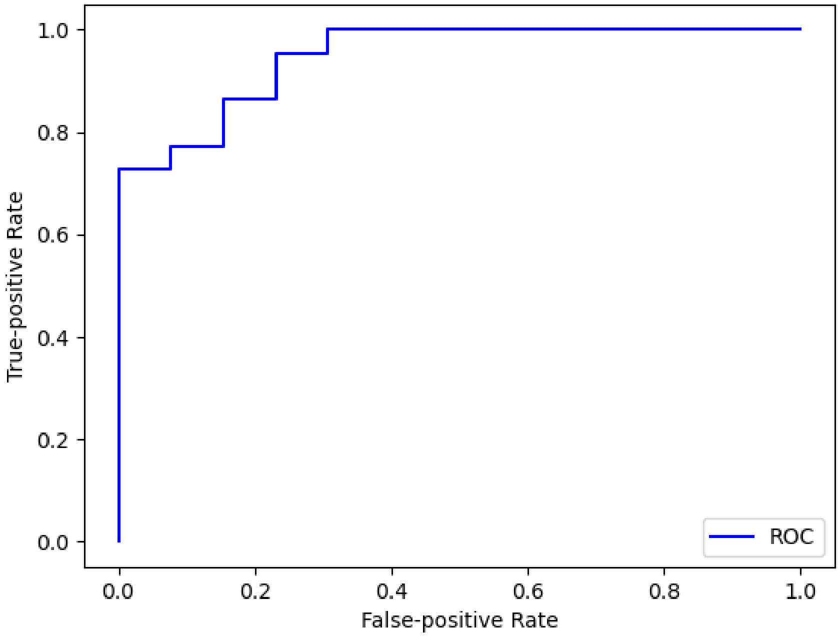

Figure 4.5: Random Forest RFECV ROC Curve

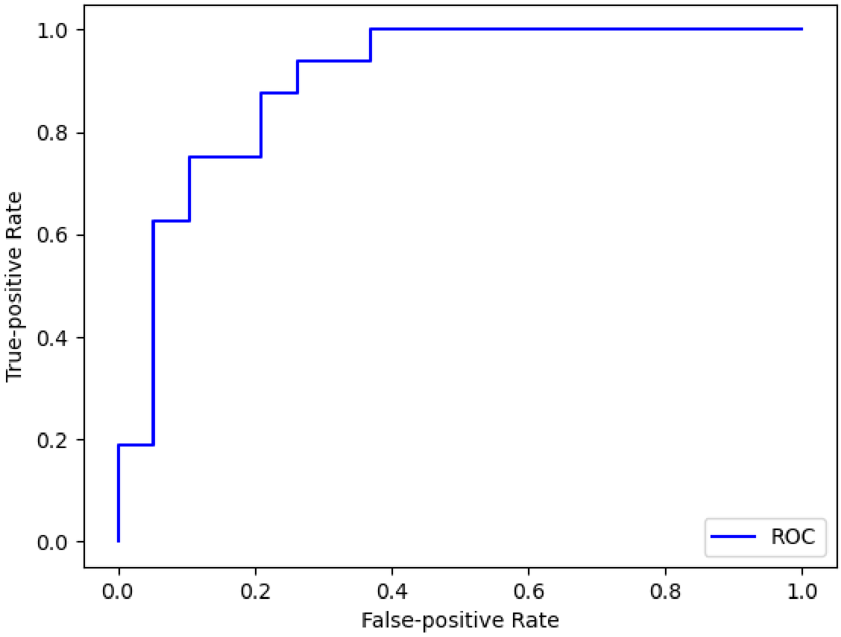

Figure 4.6: Support Vector Machine RFECV ROC Curve

As shown in figure 4.7, the confusion matrix provides a high-level summary of the results obtained during the testing phase of ada boost using previously unseen data. The horizontal axis corresponds with the labels the classifier predicted while the vertical axis corresponds with the actual or true labels as specified in the original dataset. To gain insight from the confusion matrix, a predicted label and a true label are selected then the square where both labels intersect gives insight into the performance of the adaptive boosting classifier. Since the classification task is focused on identifying loan defaulters, the label “0” denotes loan non-defaulters while the label “1” denotes loan defaulters. In the figure, when the predicted label of “0” and the true label of “0” are selected, their intersect yields the number of true negative instances, 13, as classified by ada boost. “0” is considered as negative because it is not the label of interest while “1” is considered as positive because it is the label of interest. Furthermore, the number of false-negative instances is determined by selecting the predicted label of “0” and finding the intersection with the true label of “1” which yields 5 instances. Similarly, the number of true positive instances is 15 when the predicted label of “1” is matched with the true label of “1”. Finally, the number of false-positive instances is 2 when the predicted label of “1” is matched with the true label of “0”. Overall, it is established that ada boost performs well in detecting most of the loan default instances compared to loan non-defaulters.

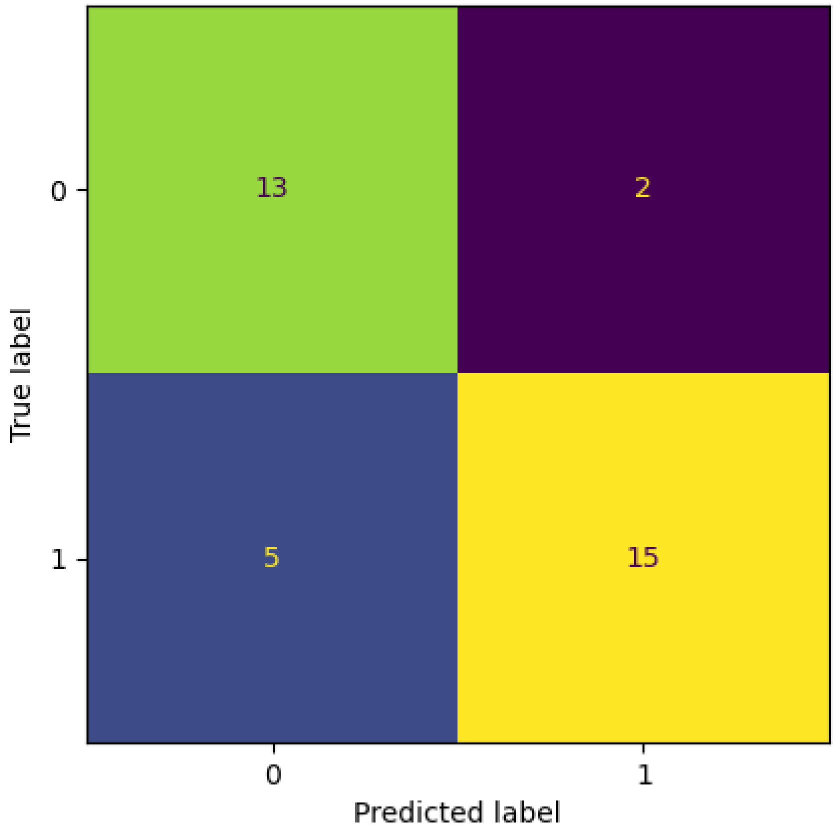

Figure 4.7: Confusion Matrix for Adaptive Boosting on RFECV Selected Features

Similarly, the result of the testing phase of gradient boosting algorithm is presented in a confusion matrix as illustrated in figure 4.8. Here, gradient boosting was able to classify 12 negative instances correctly. Consequently, only 1 negative instance was misclassified as positive. That is, only one agricultural loan non-defaulter was misclassified as a potential agricultural loan defaulter. Furthermore, 19 loan defaulters were correctly identified as the true positive instances while 3 loan non-defaulters were misclassified as potential agricultural loan defaulters. As a result, gradient boosting performs well in detecting both loan defaulters and loan non-defaulters.

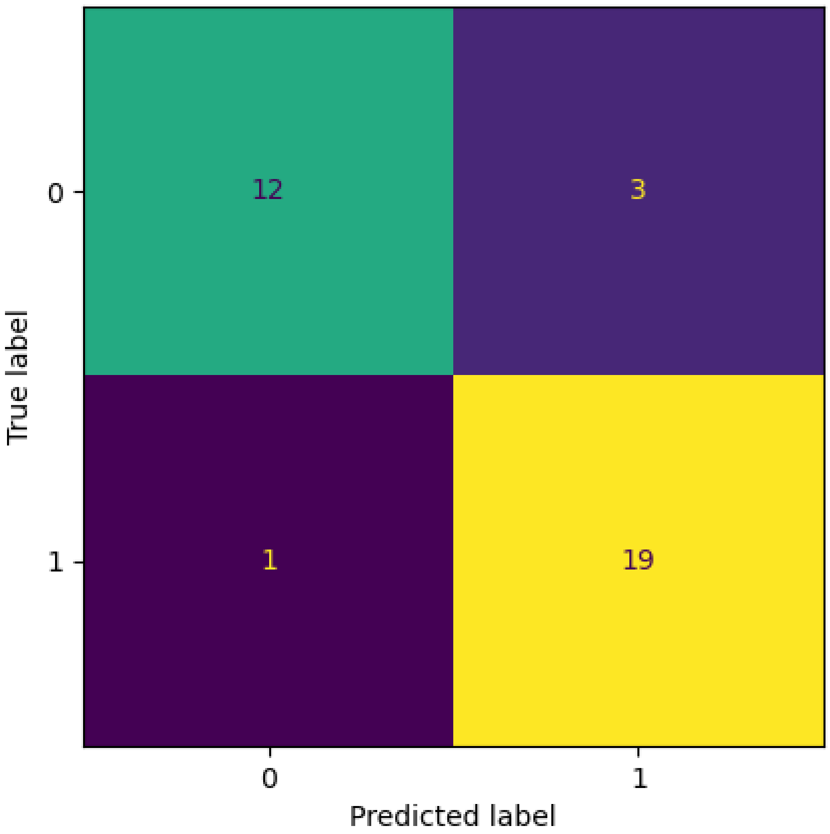

Figure 4.8: Confusion Matrix for Gradient Boosting on RFECV Selected Features

As shown in the confusion matrix in figure 4.9, logistic regression was able to identify 13 loan defaulters out of the total 14 loan defaulters present during the testing phase. The sum of loan defaulters is obtained by adding all horizontal values of the true label “1”. That is, each value in each intersect that lies on a horizontal path with the label “1” are the number of loan defaulters present in the testing dataset. Furthermore, the classifier was able to identify 16 agricultural loan non-defaulters out of a total number of 21. That is, although the logistic regression was able to identify 16 loan non-defaulters, it misclassified 5 loan non-defaulters as loan defaulters. The confusion matrix shows that logistic regression predicts loan defaulters better than loan non-defaulters. Since the interest of this research lies in predicting agricultural loan defaults, it can be said that logistic regression performs well in this area.

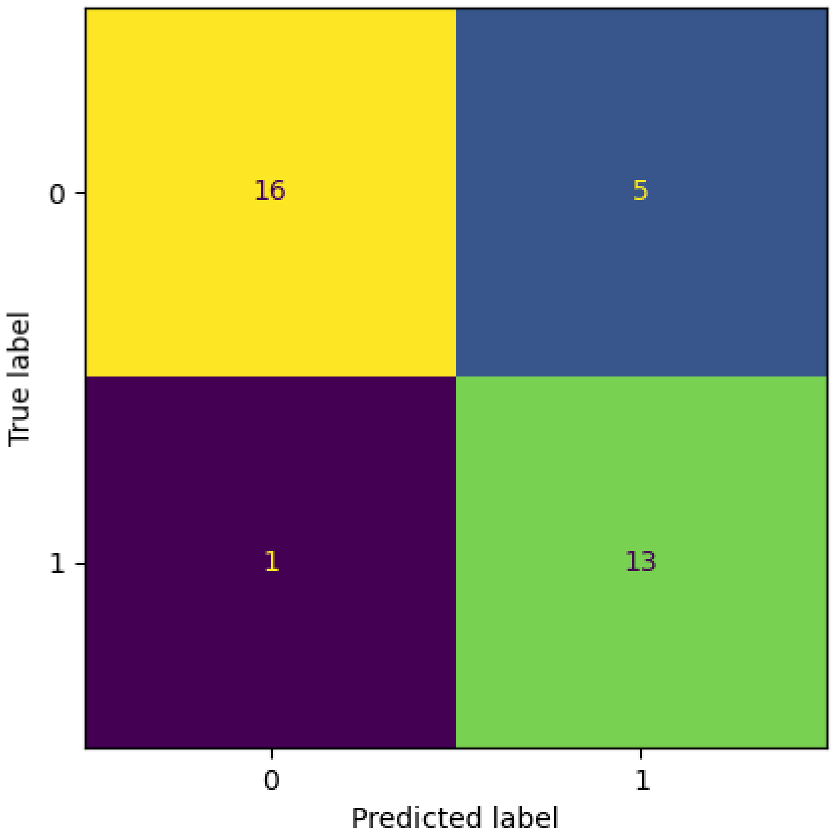

Figure 4.9: Confusion Matrix for Logistic Regression on RFECV Selected Features

The ssconfusion matrix shown in figure 4.10 illustrates the number of instances correctly classified and misclassified by random forest. Here, the random forest classifier was able to identify 17 agricultural loan defaulters but misclassified 5 as loan non-defaulters. Furthermore, the number of loan non-defaulters correctly identified is 11. The number of loan non-defaulters misclassified as defaulters is 2. Based on the results obtained, random forest is a good classifier for predicting agricultural loan defaulters but it pales in comparison to predicting non-defaulters.

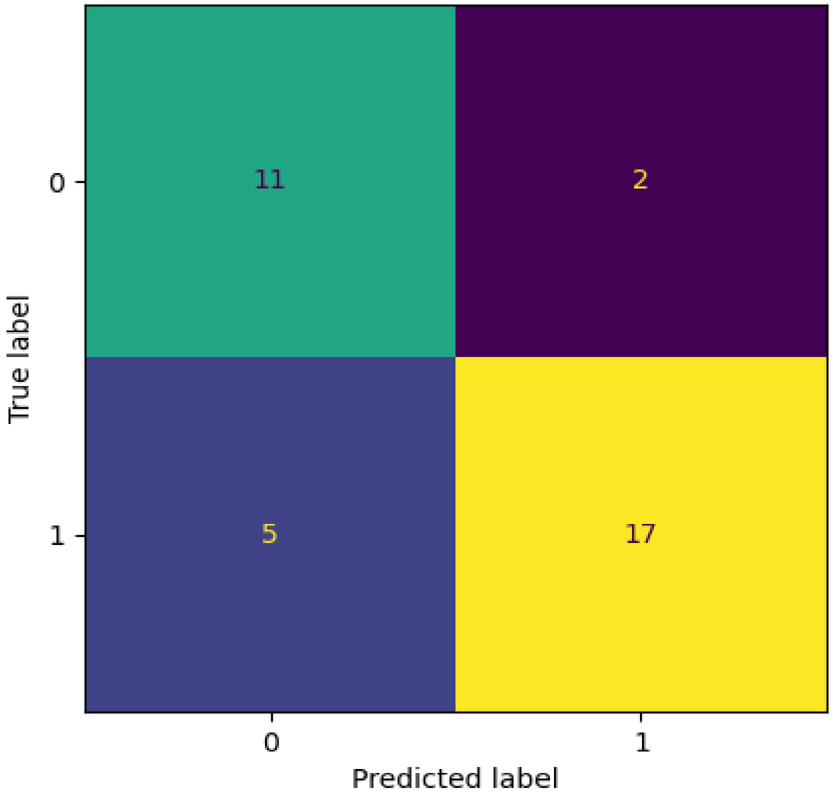

Figure 4.10: Confusion Matrix for Random Forest on RFECV Selected Features

The confusion matrix shown in fsigure 4.11 presents the number of instances correctly classified and misclassified by the support vector machine algorithm. The illustration shows that 14 instances from the testing dataset were correctly classified as loan defaulters and non-defaulters. Conversely, 5 loan non-defaulters were misclassified as loan defaulters while 2 loan defaulters were misclassified as loan non-defaulters. Consequently, support vector machine is good in detecting loan defaulters compared to loan non-defaulters.

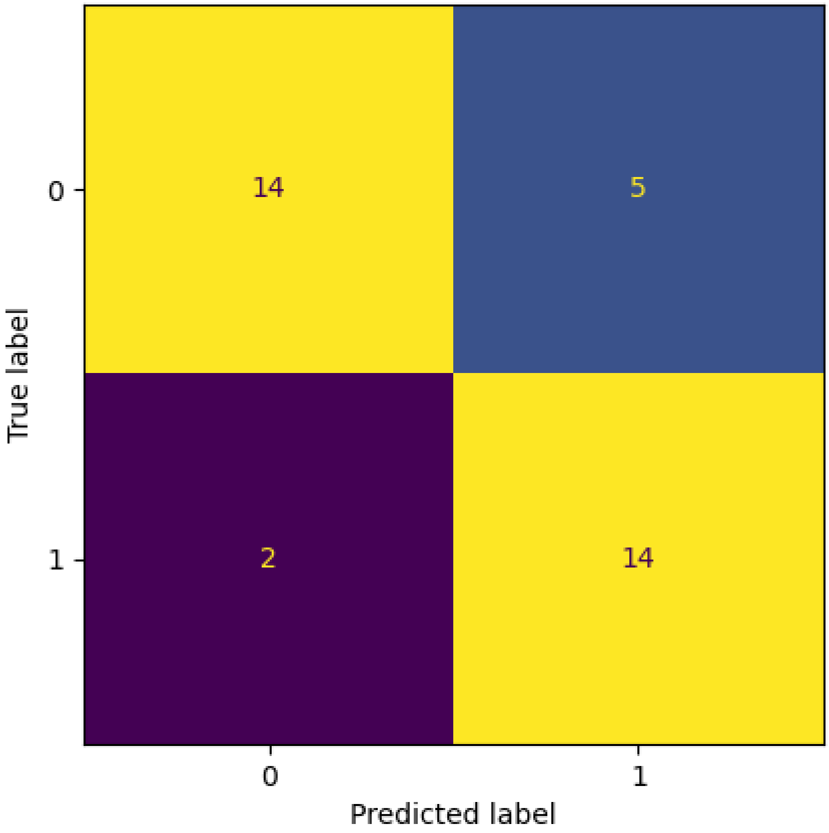

Figure 4.11: Confusion Matrix for Support Vector Machine on RFECV Selected Features

4.3 Classification of PCA Extracted Features

Similarly, the machine learning algorithms selected for the classification process were applied to the extracted features. Furthermore, the performance of the models was measured using accuracy, precision, recall, AUC, and f1-score metrics. Also, the receiver operating characteristic curve was plotted to show the performance of each model at all classification thresholds. The confusion matrix for each classifier is also presented. Table 4.3 shows the values obtained by each classifier for each performance metric while figure 4.12 illustrates the values represented in Table 4.3.

Table 4.3: **Classifier Performance Evaluation for PCA Extracted Features**

Classifiers	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1-score**
Gradient Boosting	60.00%	0.5971	56.25%	56%	56.25%
Ada Boost	74.29%	0.7435	72.22%	76.47%	74.29%
Random Forest	71.43%	0.722	65.00%	81.25%	72.22%
SVM	65.71%	0.6546	62.50%	62.50%	62.50%
Logistic Regression	74.29%	0.7417	78.95%	75.00%	76.92%

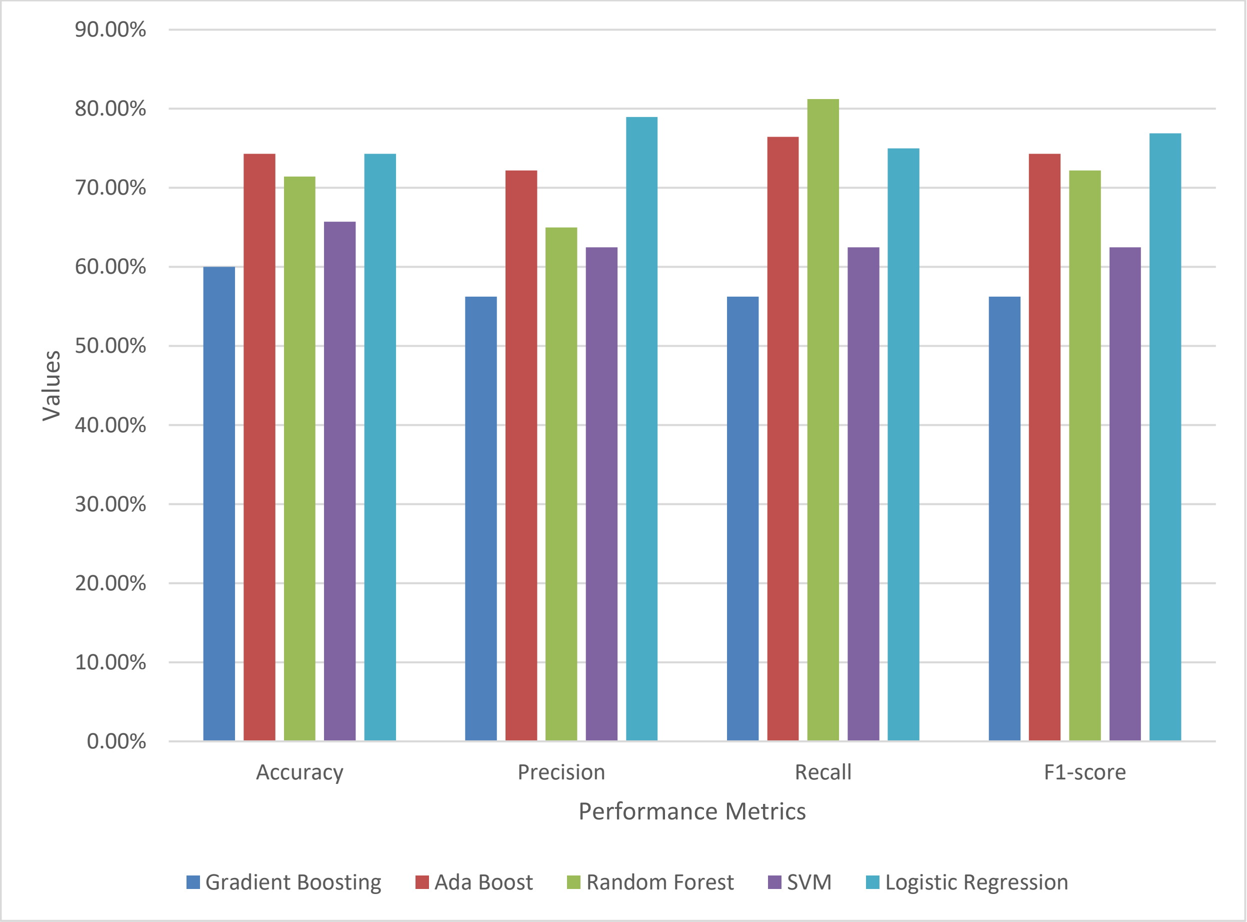

Figure 4.12: Performance of Classifiers on PCA Extracted Features

Ada Boost and logistic regression achieved the highest accuracy of 74.29% and 74.29% respectively when the classifiers were applied to the PCA extracted features.Random forest, support vector machine, and gradient boosting achieved an accuracy of 71.43%, 65.71%, and 60% respectively. Since accuracy shows how well a classifier distinguishes both loan defaulters and non-defaulters, ada boost and logistic regression have proven to be competitive.

In line with the accuracy results obtained, ada boost and logistic regression obtained the highest AUC values of 0.7435 and 0.7417 respectively. Random forest, support vector machine, and gradient boosting achieved AUC values of 0.722,0.6546, and 0.5971 respectively. The AUC values show that ada boost and logistic regression achieved better instance separability compared to the other classifiers. This means that ada boost and logistic regression were able to classify more loan defaulters as defaults and loan non-defaulters as non-default adequately than any other classifier used in this research.

For the precision metric, ada boost and logistic regression proved to be superior to other classifier by achieving precision scores of 72.22% and 78.95% respectively. This means that ada boost and logistic regression were able to identify a larger number of loan defaulters as defaults (true positive) while achieving a low false positive which classifies loan non-defaulters as defaults.The consistency of ada boost and logistic regression in achieving high scores when applied on the PCA extracted dataset shows that both classifiers are suitable in predicting agricultural loan defaults.

For the recall, random forest proved to be the most suitable in predicting agricultural loan defaults. However, ada boost and logistic regression trailed behind with recall scores of 76.47% and 75% respectively. This means that random forest, ada boost, and logistic regression were able to identify agricultural loan defaulters appropriately. However, support vector machine and gradient boosting achieved low recall scores of 62.50% and 56%. When the recall score of each classifier is juxtaposed with the precision score of the classifier, it is worthy of note that ada boost and logistic regression which both attained high precision and recall scores classify more loan defaulters as such while minimizing the number of loan non-defaulters that were misclassified as agricultural loan defaulters.

Logistic regression proves that it is the best classifier to predict loan defaulters using PCA extracted features by achieving an f1-score of 76.92%. Trailing behind is ada boost with an f1-score of 74.29%. The consistency of ada boost and logistic regression in obtaining performance values better than the other classifiers shows that both classifiers work best in predicting agricultural loan defaulters on features extracted by PCA. Random forest, support vector machine, and gradient boosting achieved f1-score of 72.22%, 62.50%, and 56.25% respectively. The recall score of random forest has shown that the classifier is also competitive in predicting agricultural loan defaulters. The ROC curves of ada boost, gradient boosting, logistic regression, random forest, and support vector machine are presented in figure 4.13, 4.14, 4.15, 4.16, and 4.17 respectively.

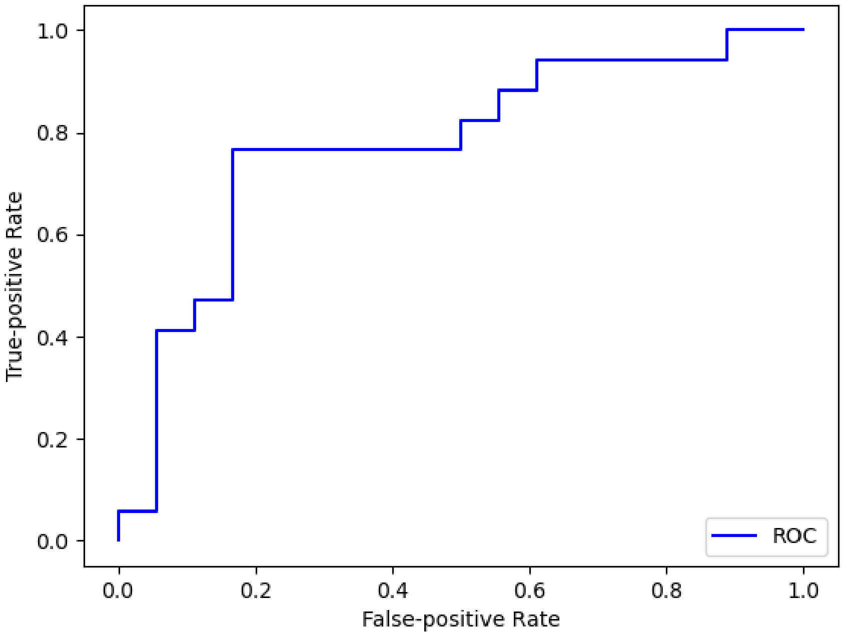

Figure 4.13: Adaptive Boosting PCA ROC Curve

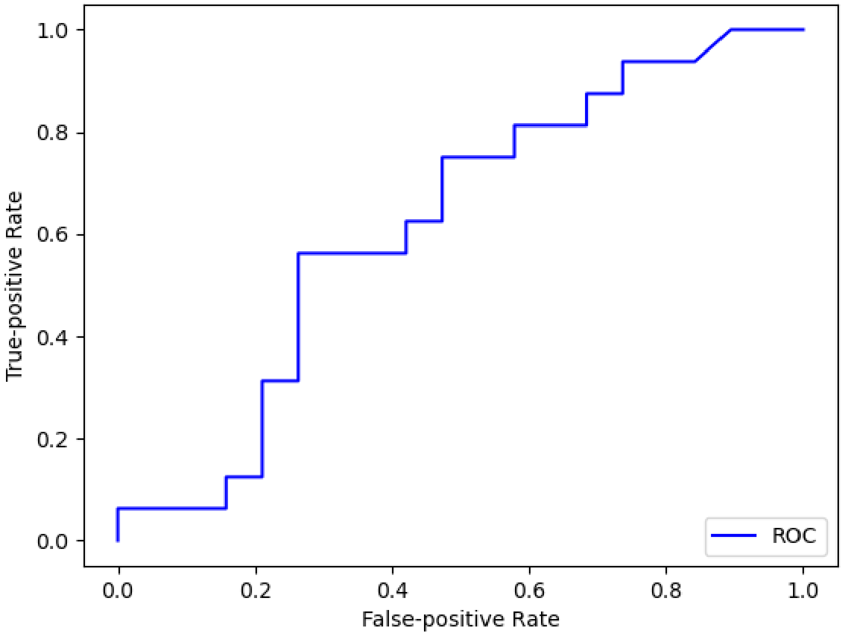

Figure 4.14: Gradient Boosting PCA ROC Curve

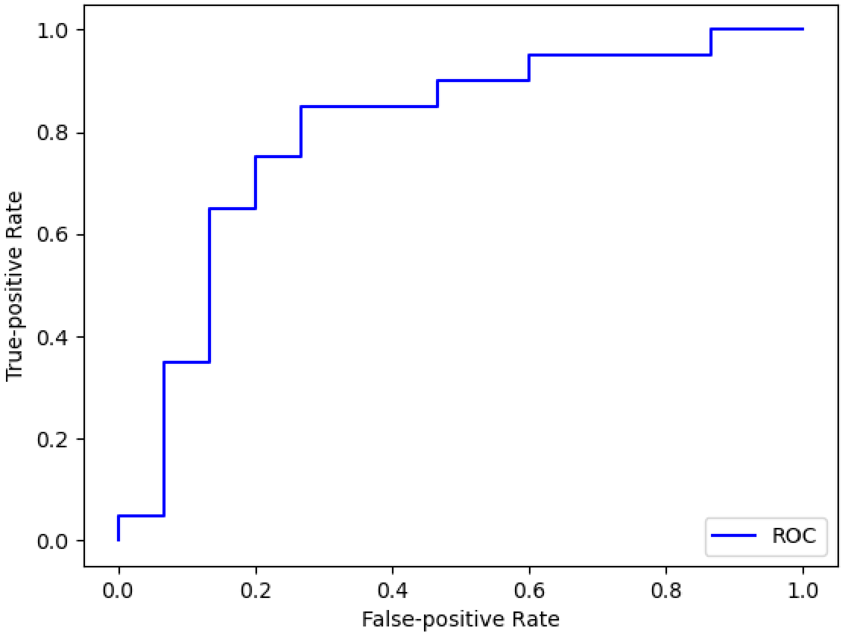

Figure 4.15: Logistic Regression PCA ROC Curve

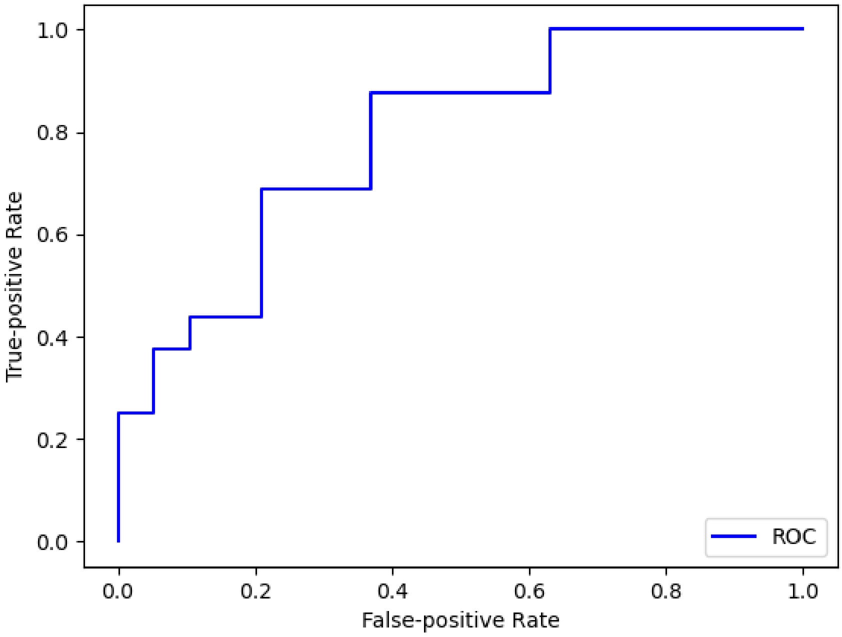

Figure 4.16: Random Forest PCA ROC Curve

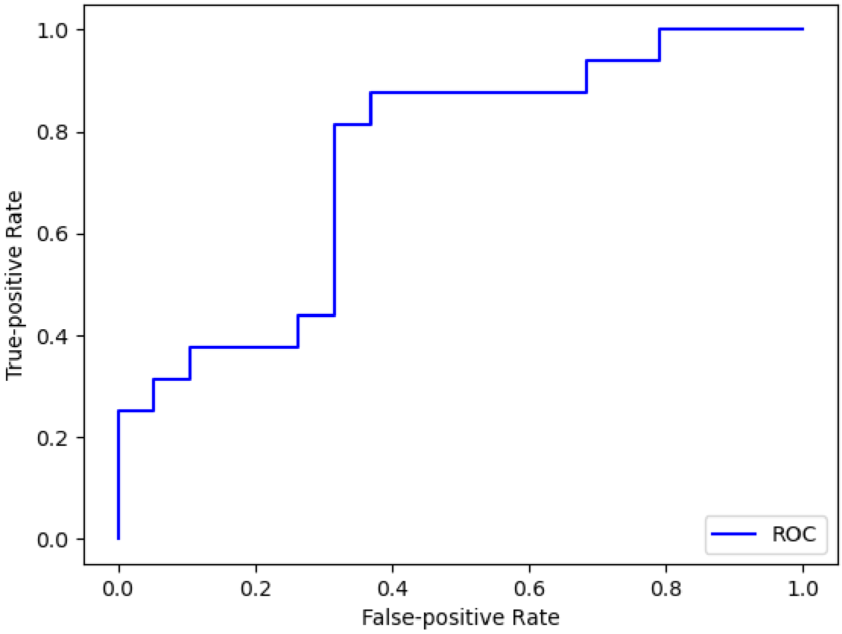

Figure 4.17: Support Vector Machine PCA ROC Curve

As shown in figure 4.18, the confusion matrix provides a visualization of the number of instances that make up the detailed results presented in table 4.3. Here, the horizontal axis corresponds with the predicted labels while the vertical axis corresponds with the actual or true labels in the original dataset. To interpret a confusion matrix, a predicted label and a true label are selected then the intersection found to obtain insight into the performance of the classifier. Since the classification task is focused on identifying loan defaulters, the label “0” denotes loan non-defaulters while the label “1” denotes loan defaulters. In the figure, ada boost had 13 true negative and positive instances. For the false negative and positive instances, ada boost falsely predicted 4 loan defaulters as non-defaulters and 5 non-defaulters as defaulters.

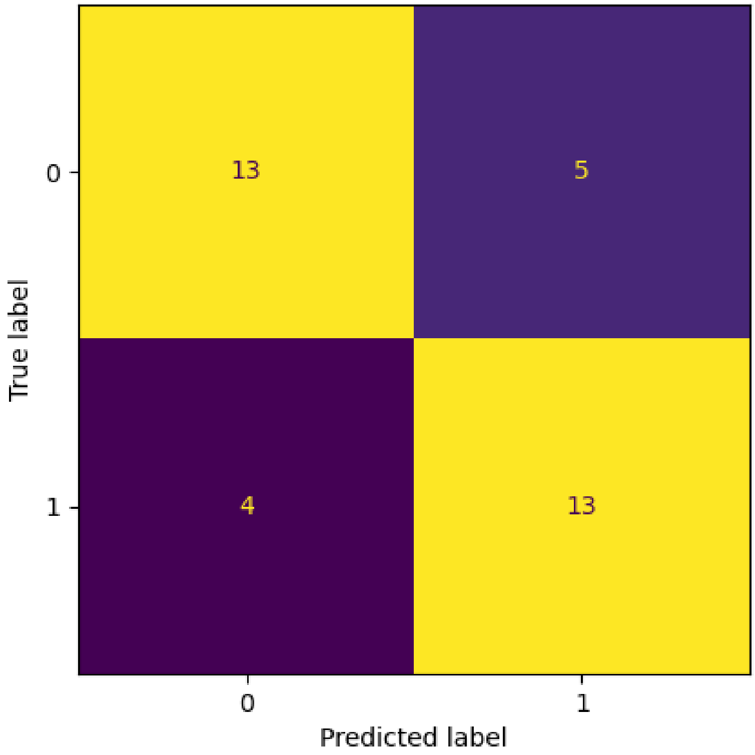

Figure 4.18: Confusion Matrix for Adaptive Boosting on PCA Extracted Features

In figure 4.19, gradient boosting identified 12 loan non-defaulters and 9 loan defaulters correctly while misclassifying 7 loan default instances as non-defaulters and another 7 loan non-default instances as defaulters. On juxtaposing the number of correctly classified loan defaulters (true positive) with the number of misclassified loan defaulters, it is established that gradient boosting struggles in predicting agricultural loan defaulters from PCA extracted features.

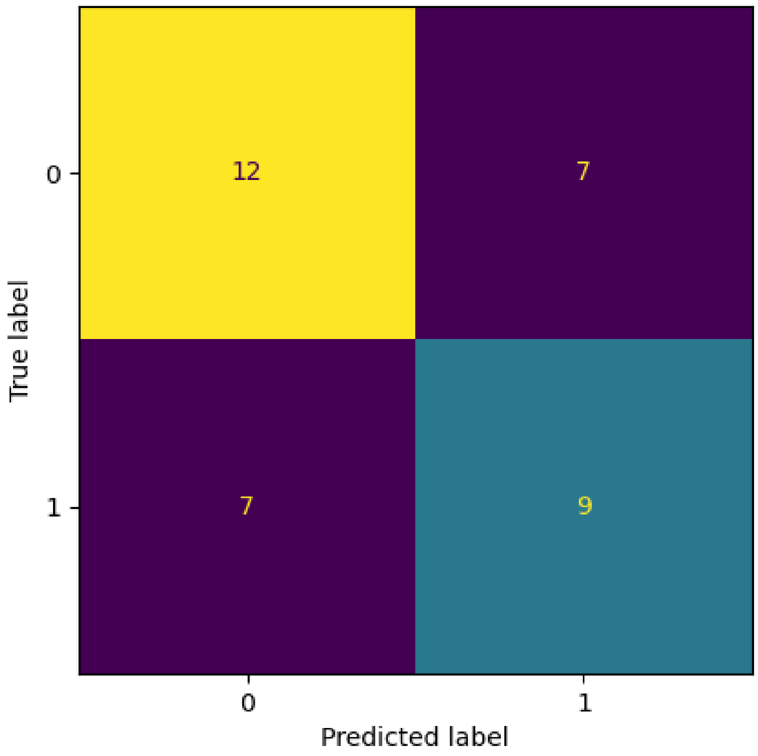

Figure 4.19: Confusion Matrix for Gradient Boosting on PCA Extracted Features

Logistic regression correctly classified 15 loan defaulters and 11 loan non-defaulters as shown in figure 4.20. However, it misclassified 4 loan non-default instances as default and 5 loan default instances as non-default.

In figure 4.21, the random forest was able to correctly classify 13 loan defaulters and 12 loan non-defaulters. However, it misclassified 3 loan defaulters and 7 loan non-defaulters.

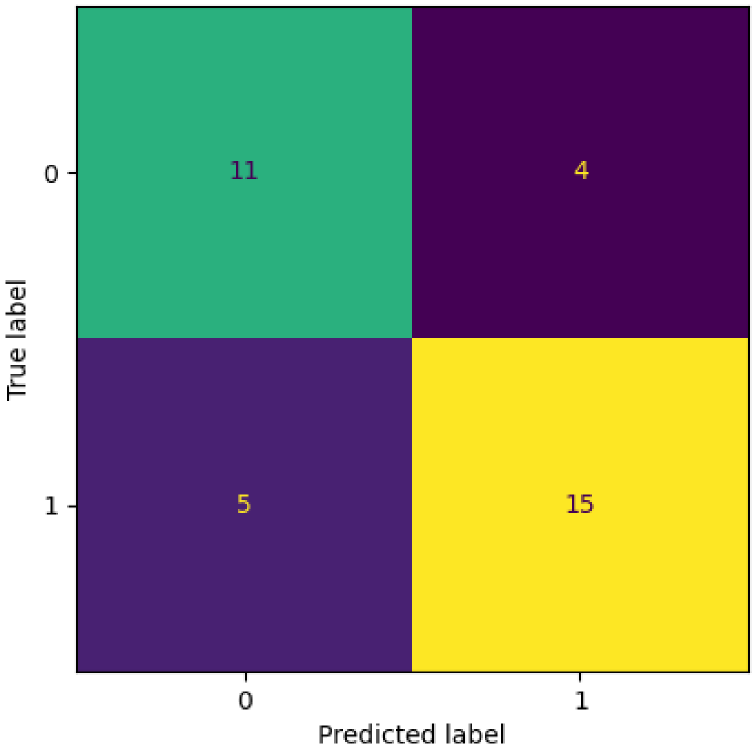

Figure 4.20: Confusion Matrix for Logistic Regression on PCA Extracted Features

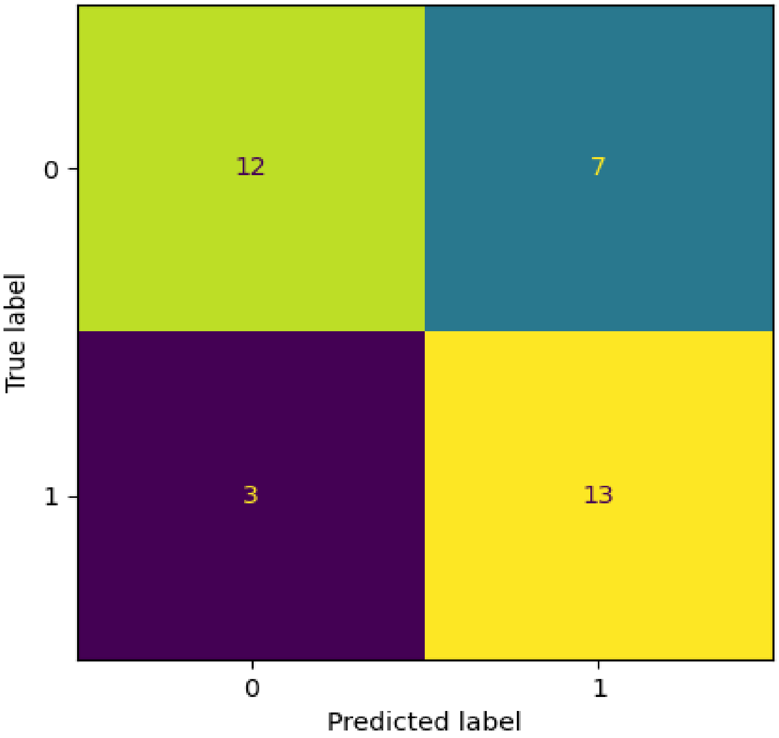

Figure 4.21: Confusion Matrix for Random Forest on PCA Extracted Features

As shown in figure 4.22, the support vector machine correctly classified 10 loan default instances and 13 loan non-default instances. However, it misclassified 6 loan default and non-default instances. That is, it classified 6 loan default instances as non-default instances and another 6 loan non-default instances as default.

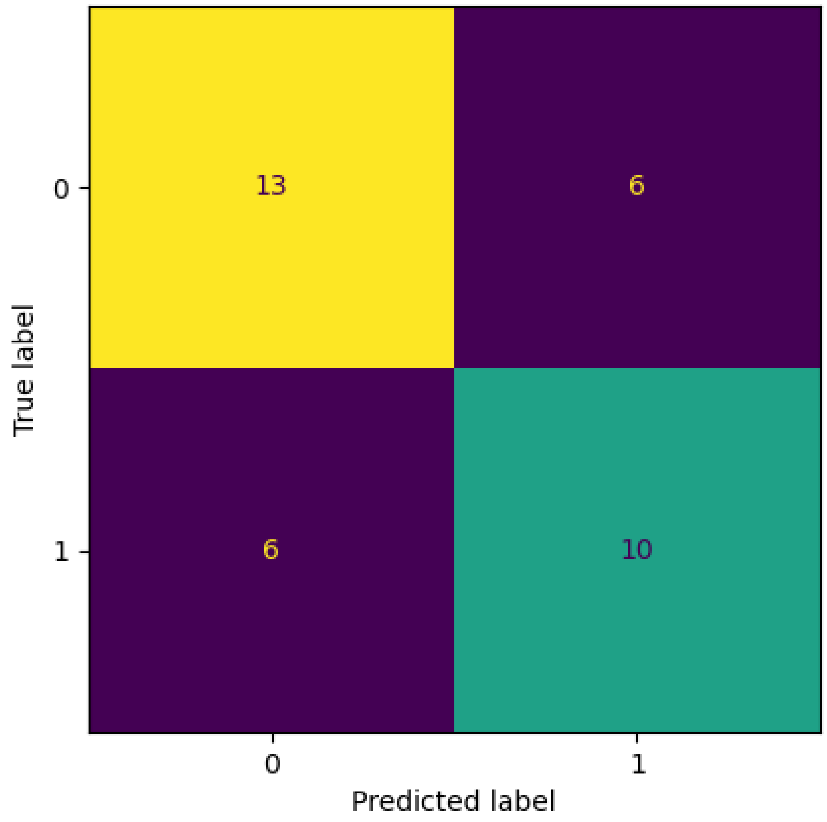

Figure 4.22: Confusion Matrix for Support Vector Machine on PCA Extracted Features

4.4 Cross-validation of RFECV Selected Features

For the RFECV selected features, the gradient boosting model is the most effective model in predicting agricultural loan defaulters among smallholder farmers in Lavun local government Niger state, Nigeria. However, due to the stochastic nature of machine algorithms, cross-validation was applied to validate the performance of the models. K-Folds cross-validation was employed with the number of splits (K) set to 10. Table 4.4 describes the accuracy obtained for each fold in the 10-fold cross-validation technique in percentage while figure 4.23 illustrates the accuracy for each fold graphically.

Table 4.4: Accuracy in percentage for each of the 10-fold Cross-validation on RFECV Selected Features

Classifier /Fold	1	2	3	4	5	6	7	8	9	10
Random Forest	83.33	83.33	61.11	72.22	58.82	94.12	94.12	82.35	58.82	82.35
SVM	88.89	61.11	61.11	72.22	64.71	82.35	94.12	70.59	64.71	70.59
Gradient Boosting	94.44	77.78	66.67	66.67	64.71	82.35	94.12	88.24	64.71	82.35
Ada Boosting	72.22	83.33	72.22	72.22	58.82	82.35	88.24	70.59	76.47	76.47
Logistic Regression	72.22	66.67	66.67	61.11	58.82	82.35	88.24	64.71	58.82	82.35

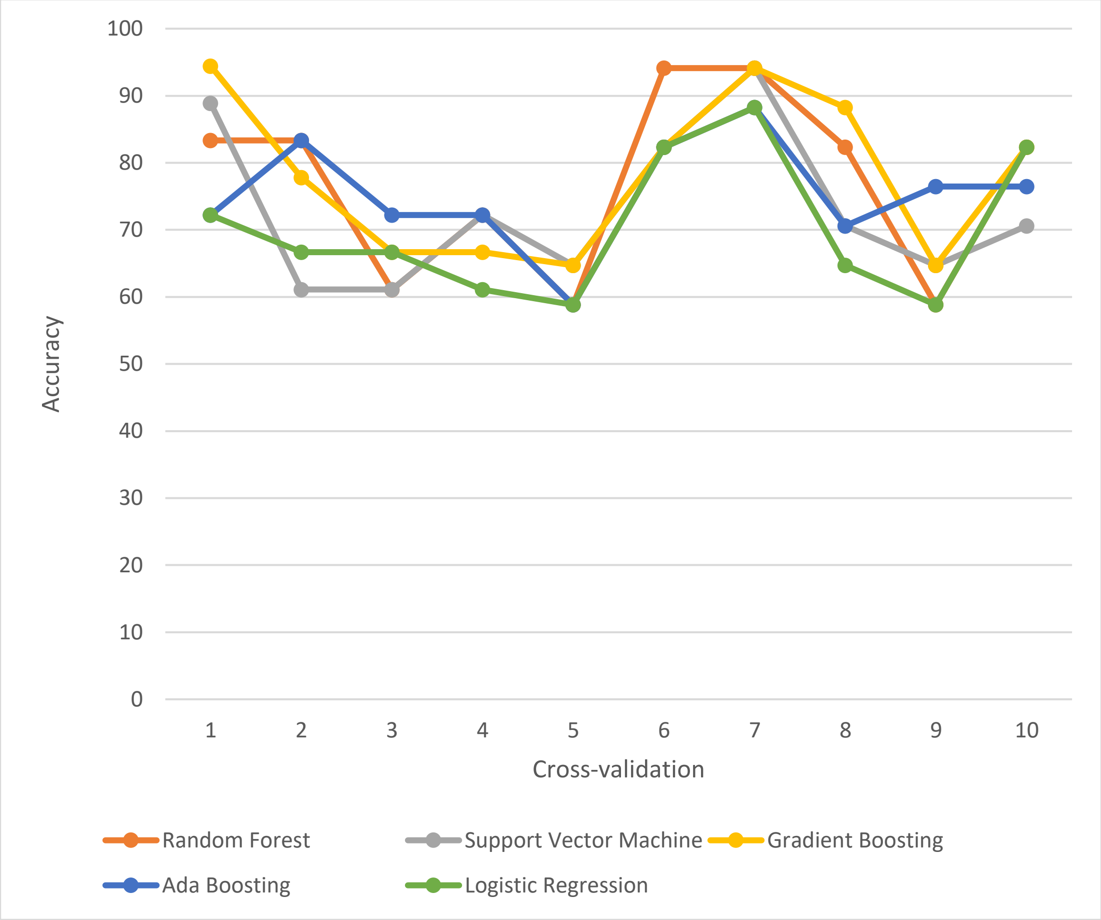

Figure 4.23: 10-fold Cross-validation Accuracy for each fold on RFECV Selected Features

From the results of the cross-validation technique, it can be noted that gradient boosting attained good accuracy for each fold in the 10-fold cross-validation. Gradient boosting even achieved the highest scores in four folds compared to Ada boost and random forest that both achieved the highest scores in five folds. To further understand the cross-validation result, the average accuracy is computed and presented in table 4.5. Figure 4.24 illustrates the average accuracy in graphical form.

Table 4.5: Average Accuracy for each Model on the 10-fold Cross-validation

Machine Learning Algorithm	Average Accuracy
Random Forest	77.06%
Support Vector Machine	73.04%
Gradient Boosting	78.20%
Ada Boosting	75.29%
Logistic Regression	70.20%

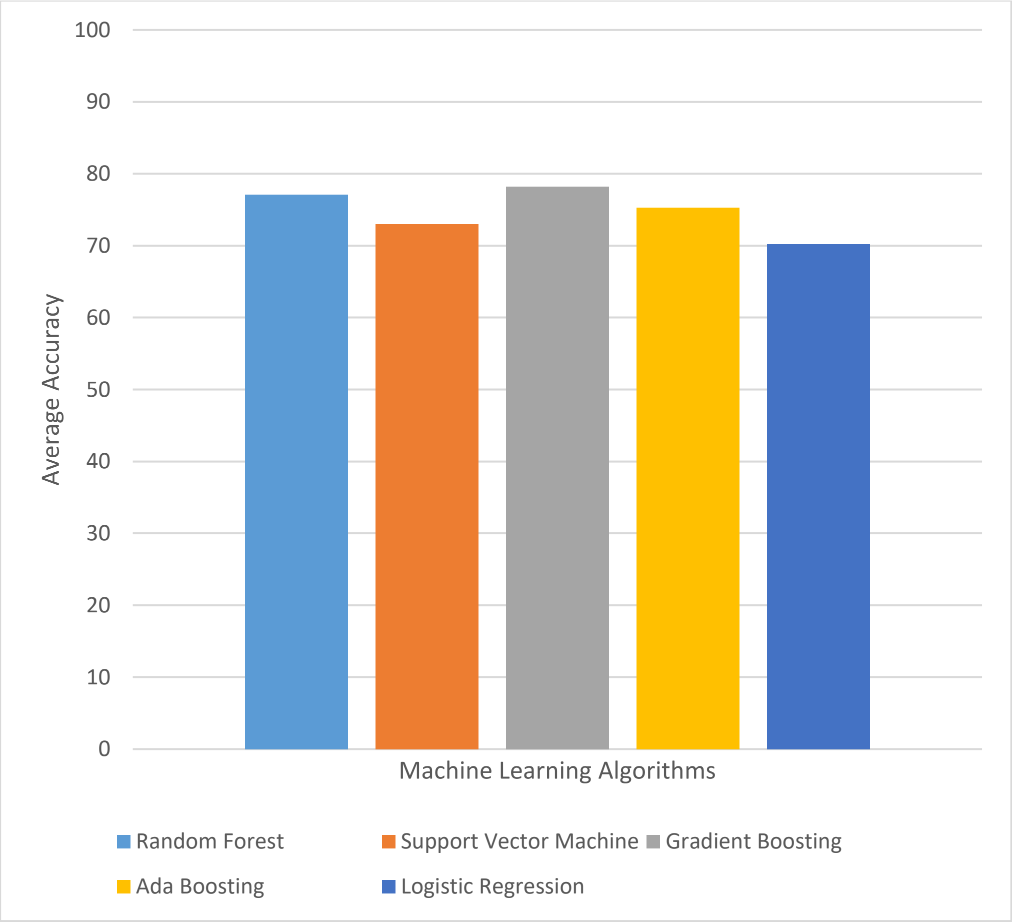

Figure 4.24: Average Accuracy for each Model on the 10-fold Cross-validation of RFECV Selected Features

The results attained when the average accuracy for each model in the 10-fold cross-validation process shows that gradient boosting had the best accuracy of 78.20%. This is followed by random forest with an average accuracy of 77.06%. The average accuracy result corroborates the earlier stated observation that gradient boosting model is the most effective model - among all the other models it was compared within this work - in predicting agricultural loan defaulters using the data gotten from smallholder farmers in Lavun local government Niger state, Nigeria. But, the ability of the gradient boosting and random forest models to produce consistent results after multiple invocations was found to be of lower consistency compared to logistic regression, ada boosting, and support vector machine; although, the low consistency rate is at an acceptable range. The ability of the algorithms to produce consistently related accuracy values was measured by taking the standard deviation of the accuracy of each model after the 10-fold cross-validation. The results of the standard deviation of each model's accuracy is shown in table 4.6 and illustrated in figure 4.25s.

Table 4.6: Standard Deviationof each Model's 10-fold Cross-validation Accuracy on RFECV Selected Features

Machine Learning Algorithm	Standard Deviation
Random Forest	0.129
Support Vector Machine	0.110
Gradient Boosting	0.113
Ada Boosting	0.078
Logistic Regression	0.101

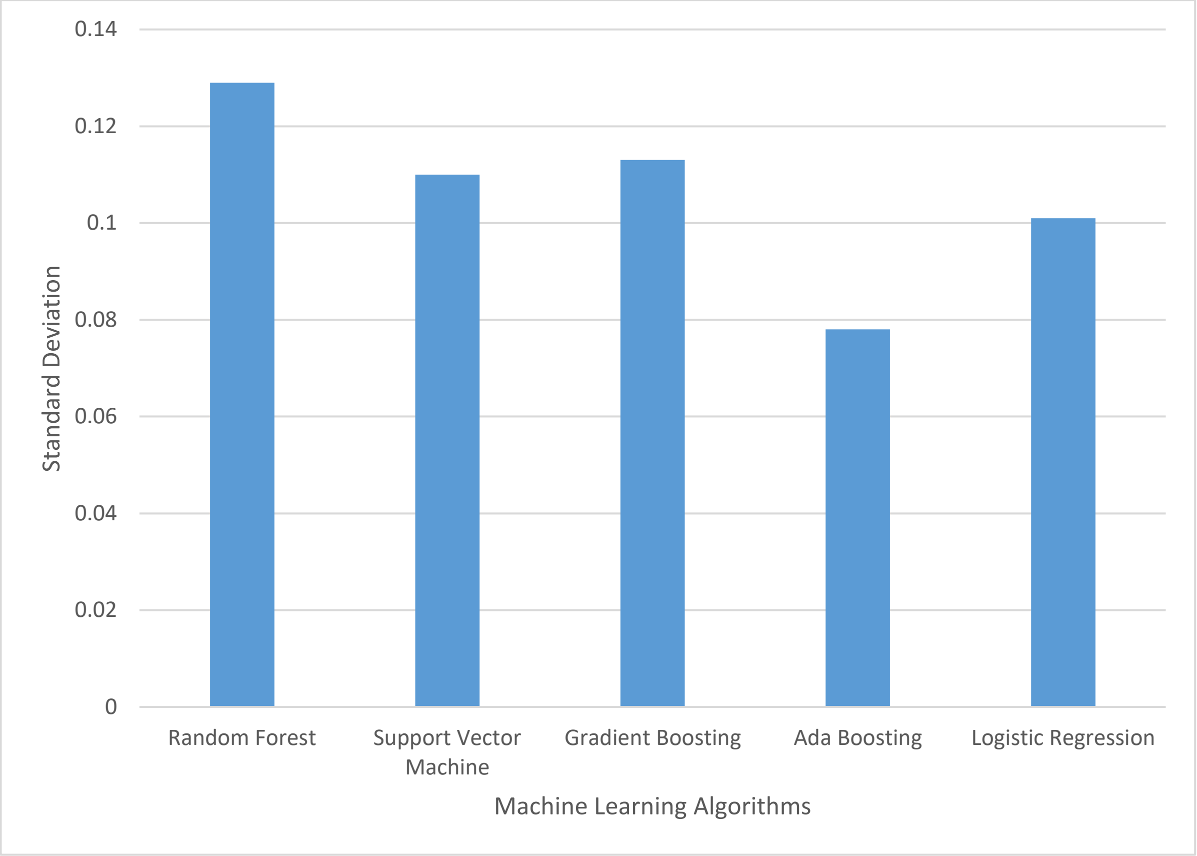

Figure 4.25: Standard Deviation of each Model’s 10-fold Cross-validation Accuracy

Hence, although gradient boosting and random forest had the best performance, the consistency of the accuracy values obtained from these models is poor compared to Ada boosting, logistic regression, and SVM. However, since the ability to predict loan defaulters accurately is of importance compared to accurately predicting non-defaulters, it holds that gradient boosting and random forest achieve acceptable results given the high f1-score both models achieved.

4.5 Cross-validation of PCA Extracted Features

Ada boosting and logistic regression has proven to work best on PCA extracted features for predicting agricultural loan default. To further validate the findings, 10-fold cross-validation was performed using each classifier on PCA extracted features. The accuracy obtained from the cross-validation process is described in Table 4.7 and illustrated in figure 4.26.

Table 4.7: Accuracy in percentage for each of the 10-fold Cross-validation on PCA Extracted Features

Classifier /Fold	1	2	3	4	5	6	7	8	9	10
Random Forest	61.11	61.11	55.56	66.67	52.94	70.59	70.59	47.06	58.82	64.71
SVM	66.67	55.56	66.67	38.89	52.94	82.35	88.24	41.18	52.94	76.47
Gradient Boosting	61.11	55.56	38.89	66.67	47.06	58.82	58.82	47.06	35.29	76.47
Ada Boosting	77.78	61.11	38.89	72.22	47.06	58.82	76.47	52.94	47.06	76.47
Logistic Regression	72.22	61.11	66.67	55.56	70.59	82.35	82.35	52.94	52.94	76.47

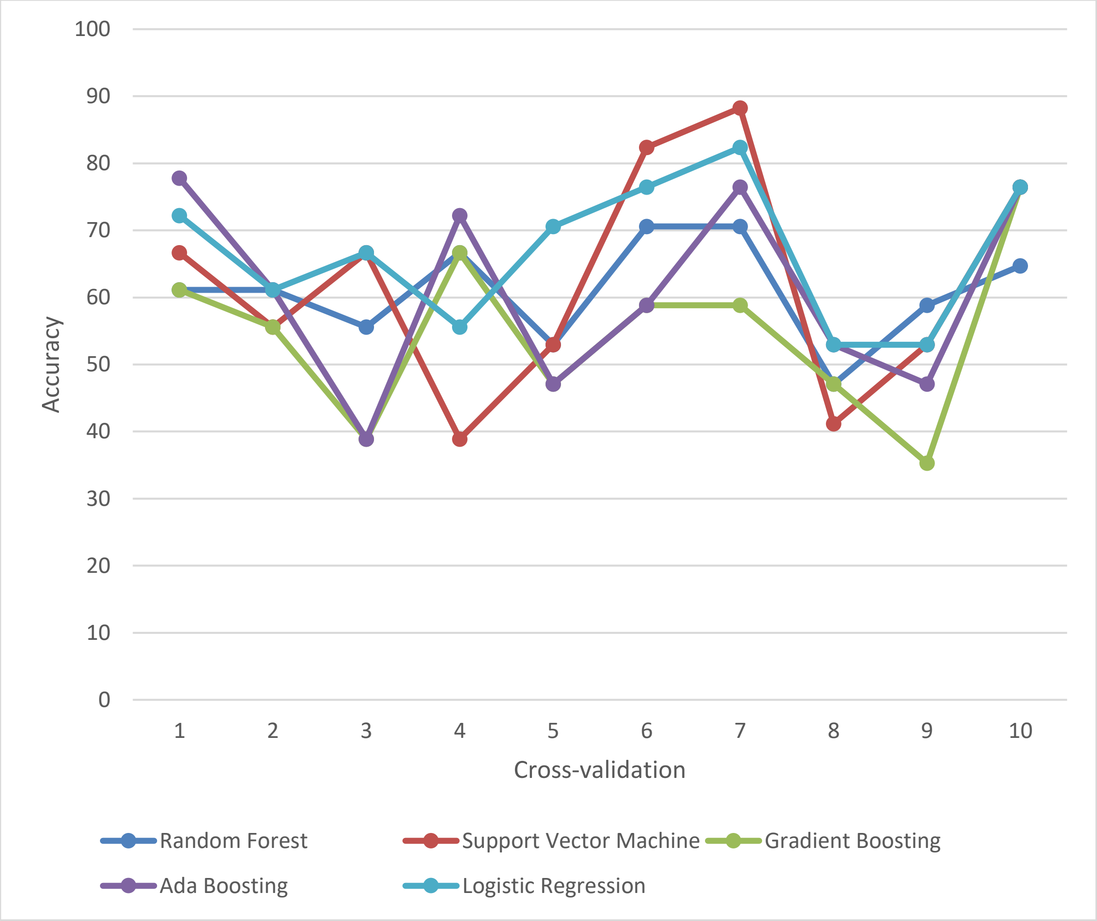

Figure 4.26: 10-fold Cross-validation Accuracy for each fold on PCA Extracted Features

As observed in the 10-fold cross-validation result in Table 4.7, ada boosting had the highest accuracy in four folds while logistic regression had the highest accuracy in six-folds. To better understand the accuracy values and draw conclusions from the cross-validation process, the average accuracy is presented in Table 4.8 and the associated illustration is given in figure 4.27.

Table 4.8: Average Accuracy for each Model on the 10-fold Cross-validation

Machine Learning Algorithm	Average Accuracy
Random Forest	60.92%
Support Vector Machine	62.19%
Gradient Boosting	54.58%
Ada Boosting	60.88%
Logistic Regression	66.73%

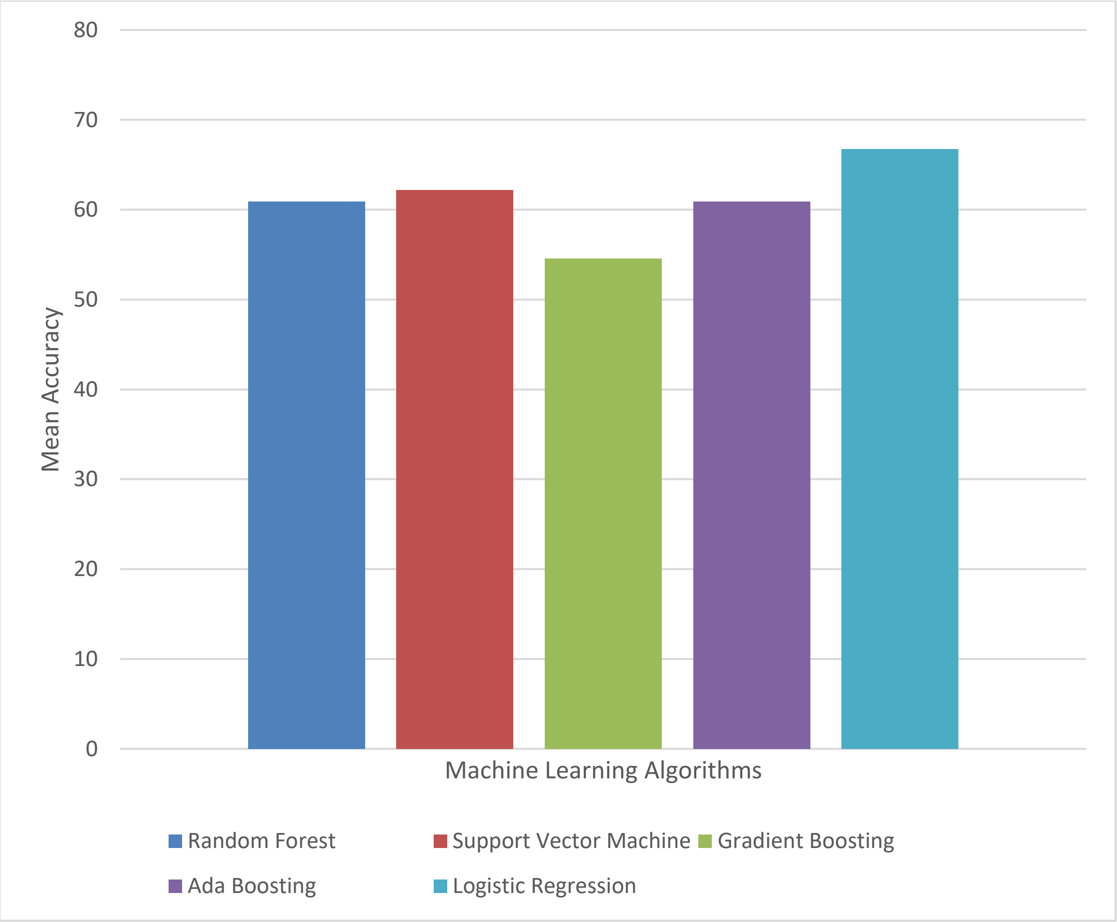

Figure 4.27: Average Accuracy for each Model on the 10-fold Cross-validation of PCA Extracted Features

The result obtained in Table 4.8 shows that logistic regression and support vector machine achieved the highest average accuracy values of 66.73% and 62.19% respectively. This is closely followed by random forest and ada boosting which both obtained average accuracy values of 60.92% and 60.88% respectively. Although support vector machine was not competitive enough in the performance metrics evaluated in Table 4.3, it was able to achieve competitive result during cross-validation. To understand how support vector machine was able to gain such a result, the standard deviation is computed. Evaluating the standard deviation of each classifier indicates how well the classifier yields consistent accuracy results. The standard deviation of each classifiers is shown in Table 4.9 and illustrated in figure 4.28.

Machine Learning Algorithm	**Standard Deviation**
Random Forest	0.072
Support Vector Machine	0.160
Gradient Boosting	0.119
Ada Boosting	0.135
Logistic Regression	0.101

Table 4.9: Standard Deviationof each Model’s 10-fold Cross-validation Accuracy on PCA Extracted Features

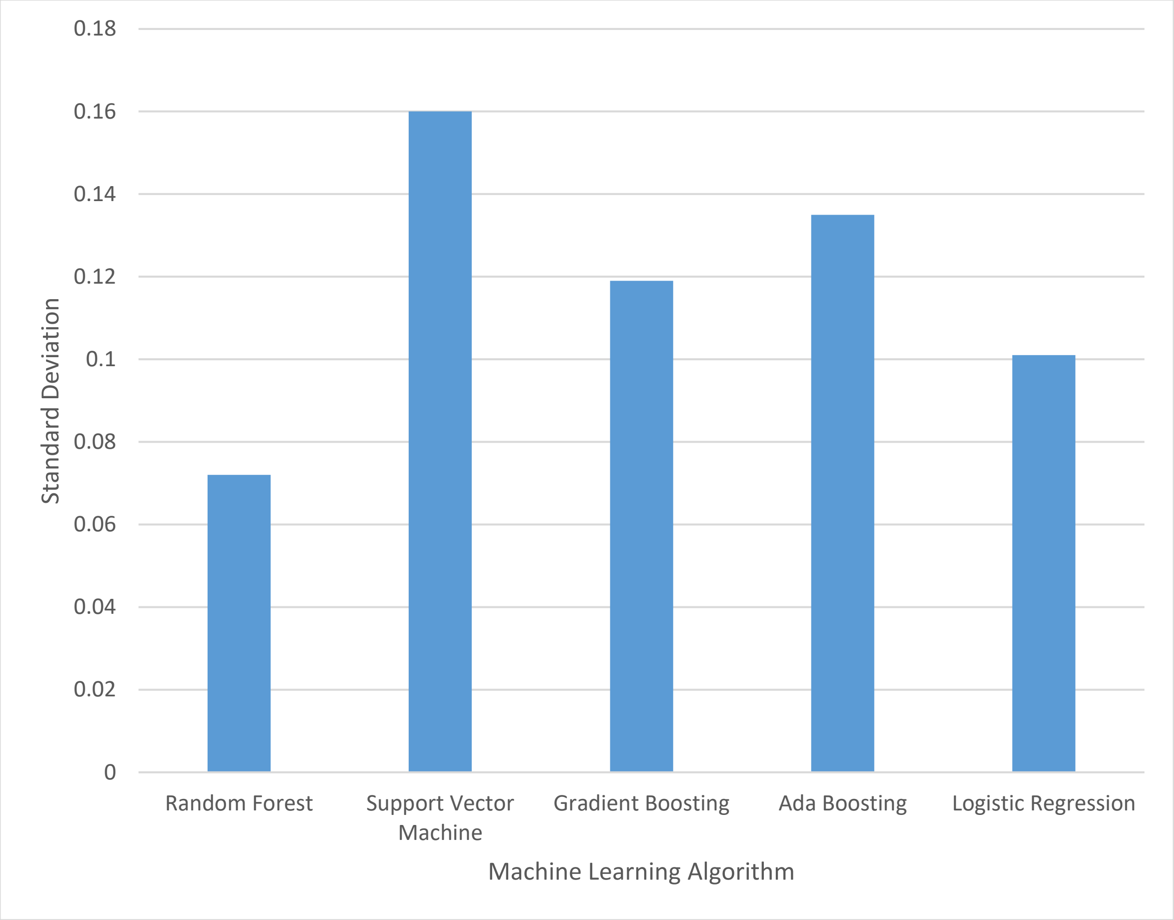

Figure 4.28: Standard Deviation of each Model’s 10-fold Cross-validation Accuracy

As evident in Table 4.9, support vector machine had the highest standard deviation. This means that compared to other classifiers, the accuracy values produced by support vector machine varied more. This contributed to the high average accuracy value obtained by support vector machine though it could not obtain competitive results as shown in Table 4.3. Random forest and logistic regression had the lowest standard deviation of 0.072 and 0.101 respectively. This means that among the other classifiers, random forest and logistic regression had the most consistent accuracy values across folds defined by the cross-validation process. Furthermore, given that logistic regression had been competitive in all metrics measured in Table 4.3, 4.8, and 4.9, it is established that logistic regression works best on predicting agricultural loan defaults using PCA extracted features.

4.6 Comparison of Results

The results obtained when features were selected using RFECV and when features were extracted using PCA are compared to gain insight into which dimensionality reduction technique works best with the classifiers in predicting agricultural loan defaults and which classifier yields the best performance metric of interest.

4.6.1 Comparison of classification using RFECV selected features and PCA extracted features

The results obtained when the classifiers were applied to the features selected by RFECV and the features extracted by PCAare juxtaposed in Table 4.10 and conclusions are drawn for each classifier and feature dimension reduction technique.

Table 4.10: Comparison of Classifiers with Dimensionality Reduction Techniques based on Performance Metrics

Classifiers + Dimensionality Reduction	Accuracy	AUC	Precision	Recall	F1-score
Gradient Boosting + RFECV	88.57%	87.5%	86.36%	95%	90.48%
Gradient Boosting + PCA	60.00%	59.71%	56.25%	56%	56.25%
Ada Boost + RFECV	80%	80.83%	88.24%	75%	81.08%
Ada boost + PCA	74.29%	74.35%	72.22%	76.47%	74.29%
Random Forest + RFECV	80%	80.94%	89.47%	77.27%	82.93%
Random Forest + PCA	71.43%	72.20%	65.00%	81.25%	72.22%
SVM + RFECV	80%	80.59%	73.68%	87.5%	79.99%
SVM+PCA	65.71%	65.46%	62.50%	62.50%	62.50%
Logistic Regression + RFECV	82%	84.52%	72.22%	92.86%	81.25%
Logistic Regression + PCA	74.29%	74.17%	78.95%	75.00%	76.92%

As shown in Table 4.10, gradient boosting with RFECV supersedes gradient boosting with PCA in all performance metrics. Furthermore, it is observed that the difference in performance metric values when RFECV and PCA features were used by the gradient boosting classifierranges from 27.79% to 69%. The large differing performance metric values shows that applying gradient boosting algorithm to features selected by RFECV yields better result in all metrics compared to when gradient boosting is applied to PCA extracted features from the agricultural loan dataset.

Similarly, when ada boost was applied to RFECV selected features, the classifier obtained better result in the performance metrics compared to when the classifier was applied to features extracted by PCA. However, ada boost with PCA achieved better recall values than ada boost with RFECV. The difference in recall values is 1.47%. This means that ada boost with PCA was able to predict agricultural loan defaulters better than ada boost with RFECV. The difference in performance metric values between ada boost with PCA and RFECV ranges from 1.47% to 16.02%. Although the differences observed is significant, it is better compared to the differences obtained when gradient boosting with RFECV was juxtaposed with gradient boosting with PCA. Therefore, ada boost with PCA is competitive in obtaining good values as well as ada boost with RFECV.

Furthermore, when random forest was applied to RFECV selected features, the result obtained was higher in most performance metrics compared to the results when random forest was applied to PCA extracted features. Regarding the difference in performance metric values based on the dimensionality reduction technique applied, the difference ranged from 8.57% to 24.47%. However, similar to the result obtained in the recall by the ada boost classifier, the recall score of random forest with PCA exceeds the recall score of random forest with RFECV.The difference in recall value is 3.98%. This means that random forest with PCA predicts agricultural loan defaulters better than random forest with RFECV. Considering the performance metrics in whole, random forest with RFECV outperforms random forest with PCA.

For the support vector machine classifier, the use of RFECV selected features ensured better performance compared to the use of PCA extracted features.The difference in performance values ranges from 11.18% to 25%. Therefore, support vector machine with RFECV selected features works best in predicting agricultural loan defaulters compared to using PCA extracted features.

In using logistic regression to predict agricultural loan defaulters in Lavun local government area of Niger state, it was established that the use of RFECV selected features produced a better result in most performance metrics compared to using PCA extracted features. Although logistic regression with REFCV outperformed logistic regression with PCA, logistic regression with PCA was able to outperform logistic regression with RFECV with 6.73% on the precision score. This means that logistic regression with PCA was able to detect more loan defaulters while minimizing the number of loan non-default misclassification compared to logistic regression with RFECV.

Overall, classifiers that used RFECV selected features fared better compared to classifiers that used the features extracted by PCA. The standard deviation for each classifier that applied each dimensionality reduction is evaluated next. This is crucial to select the classifier that produces consistent output. Table 4.11 juxtaposes the standard deviation of each classifier with each dimensionality reduction method.

Table 4.11: Comparison of Classifiers with Dimensionality Reduction Technique based on Standard Deviation

Classifier + Dimensionality Reduction	Standard Deviation
Gradient Boosting + RFECV	0.113
Gradient Boosting + PCA	0.119
Ada Boost + RFECV	0.078
Ada boost + PCA	0.135
Random Forest + RFECV	0.129
Random Forest + PCA	0.072
SVM + RFECV	0.110
SVM+PCA	0.160
Logistic Regression + RFECV	0.101
Logistic Regression + PCA	0.101

As shown in Table 4.11, gradient boosting with RFECV had consistent result compared to gradient boosting with PCA. This is evident in the high standard deviation of gradient boosting with PCA compared to gradient boosting with RFECV. Therefore, gradient boosting with RFECV yields consistent accuracy result compared to PCA. Similarly, ada boosting with RFECV yields consistent results than ada boosting with PCA. The standard deviation difference of 0.057 shows that ada boosting with PCA is not as consistent in predicting agricultural loan defaulters and loan non-defaulters. Conversely, random forest with PCA produces consistent accuracy results compared to random forest with RFECV. For support vector machine, the use of RFECV selected features yielded consistent accuracy results compared to PCA extracted features. Logistic regression is the only classifier that produced the same consistent accuracy result when both feature dimensionality techniques were used.

4.7 Gradient Boosting as the Best Machine Learning Algorithm for Predicting Agricultural Loan Defaulters

Given that features selected by RFECV have been shown to improve loan defaulter prediction compared to features extracted by PCA, the machine learning algorithm that best classifies RFECV selected features is gradient boosting.Gradient boosting produces a prediction model in a stage-wise manner by forming an ensemble of various weak predictors which are usually decision trees. It tries to improve each predictor by reducing the errors in its predecessor in relation to the actual values in the training dataset.

Furthermore, gradient boosting with RFECV supersedes gradient boosting with PCA in all performance metrics. Also, the difference observed when RFECV and PCA features were used by gradient boosting algorithm ranges higher than other algorithms. Therefore, the large differing performance metric values shows that applying gradient boosting algorithm to features selected by RFECV yields better result in all metrics compared to when gradient boosting is applied to PCA extracted features from the agricultural loan dataset.

CHAPTER FIVE

5.0 CONCLUSION AND RECOMMENDATIONS

5.1 Summary

This research compared various machine learning classifiers in predicting agricultural loan exploring defaulters among farmers in Lavun local government area of Niger state while two-dimensionality reduction techniques, recursive feature elimination with cross-validation and principal component analysis because the dataset contained 60 features and 174 unique instances.The dataset was normalised using z-score normalisation to mitigate value sensitivity which PCA and support vector machine are prone to.Then RFECV and PCA were applied to the dataset independently to select the appropriate features and extract the appropriate features respectively.The selected and extracted features were then classified using gradient boosting, ada boost, random forest, support vector machine, and logistic regression. The performance metrics used to measure the performance of each classifier are accuracy, precision, recall, AUC, and f1-score. To further establish the ability of each classifier to produce consistent result, 10-fold cross-validation was applied and the accuracy, average accuracy, and standard deviation were recorded. Furthermore, the classifiers were compared on each performance metric and dimensionality reduction technique.

5.2 Conclusion

In conclusion, it has been established that features selected by RFECV enabled the classifiers to perform better than the extracted features of PCAdue to the differences observed in section 4.6.1. On using RFECV to select features, random forest and gradient boosting performed best compared to the remaining three classifiers. On using PCA to extract features, logistic regression and ada boost performed best in predicting loan defaulters compared to the other classifiers. Overall, logistic regression proved to be superior to other classifiers whether features were selected using RFECV or extracted using PCA due to the ability of logistic regression to maintain low consistent standard deviation values of 0.101 across both dimensionality reduction techniques compared to other classifiers. Although the performance of logistic regression trailed the performance of random forest and gradient boosting when RFECV was used to select the features, it proved to be competitive enough across all the performance metrics.

5.3 Recommendations

This study applied machine learning classifiers to agricultural loan dataset obtained from Lavunwhile using features selected by RFECV and features extracted by PCA.It is recommended that further research into achieving lower standard deviation of classification accuracy when both dimensionality reduction techniques of RFECV and PCA are applied to features be explored.

5.4 Contribution to Knowledge

This research has been able to establish that logistic regression is the most stable classifier when RFECV or PCA dimensionality reduction techniques is applied to the agricultural loan dataset used in this work.Furthermore, this will aid financial institutions in minimizing risk when issuing loans to farmers.

REFERENCES

Adenekan, M. O., & Augustus, E. O. (2021). Agricultural Transformationin Nigeriafor Sustainable Food Security. *Journal of Global Biosciences*, *10*(1), 8230–8242.

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013,* India, 203–207. https://doi.org/10.1109/ICMIRA.2013.45

Al-Qerem, A., Al-Naymat, G., & Alhasan, M. (2019). Loan default prediction model improvement through comprehensive preprocessing and features selection. *Proceedings - 2019 International Arab Conference on Information Technology, ACIT 2019,* 235–240. https://doi.org/10.1109/ACIT47987.2019.8991084

Aslam, M., Kumar, S., & Sorooshian, S. (2020). Predicting likelihood for loan default among bank borrowers. *International Journal of Financial Research*, *11*(1), 318–328. https://doi.org/10.5430/ijfr.v11n1p318

Bayraci, S., & Susuz, O. (2019). A Deep Neural Network (DNN) based classification model in application to loan default prediction. *Theoretical and Applied Economics*, *XXVI*(4), 75–84.

Coşer, A., Maer-Matei, M. M., & Albu, C. (2019). Predictive models for loan default risk assessment. *Economic Computation and Economic Cybernetics Studies and Research*, *53*(2), 149–165. https://doi.org/10.24818/18423264/53.2.19.09

Datkhile, A., Chandak, K., Bhandari, S., Gajare, H., & Karyakarte, M. (2020). Statistical Modelling on Loan Default Prediction Using Different Models. *IJRESM, 3*(3), 3–5.

Deng, T. (2019). Study of the prediction of micro-loan default based on logit model. *Proceedings - 2019 International Conference on Economic Management and Model Engineering, ICEMME 2019,* Malacca, Malaysia. 260–264. https://doi.org/10.1109/ICEMME49371.2019.00058

Dushimimana, B., Wambui, Y., Lubega, T., & McSharry, P. E. (2020). Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans. *Journal of Risk and Financial Management*, *13*(8), 180. https://doi.org/10.3390/jrfm13080180

Echebiri, R. N., & Onu, D. O. (2019). Risk management strategies among smallholder arable crop farmers in Ibiono Ibom Local Governemet Area, Akwaibom State, Nigeria. *Nigeria Agricultural Journal*, *50*(1), 22–29. Retrieved from https://0-search.ebscohost.com.ujlink.uj.ac.za/login.aspx?direct=true&db=awn&AN=naj-190643&site=ehost-live&scope=site

Elakkiya, E., Radhaiah, K., & Rayalu, G. M. (2020). *Logistic regression models for prediction loan defaults-qualtitative data analysis*. *9*(8), 6027–6034.

Evbuomwan, G. O., & Okoye, L. U. (2017). *Evaluating the Prospects of the Anchor Borrowers ' Programme for Small Scale Farmers in Nigeria*. *2*(1), 1–10.

Fu, X., Ouyang, T., Chen, J., & Luo, X. (2020). Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks. *Information Processing and Management, 57*(4), 102236. https://doi.org/10.1016/j.ipm.2020.102236

Jaadi, Z. (2021). A Step-by-Step Explanation of Principal Component Analysis (PCA) | Built In. Retrieved April 27, 2021, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Kim, A., & Cho, S. B. (2019). An ensemble semi-supervised learning method for predicting defaults in social lending. *Engineering Applications of Artificial Intelligence*, *81*, 193–199. https://doi.org/10.1016/j.engappai.2019.02.014

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors (Switzerland)*, (18), 1–29. https://doi.org/10.3390/s18082674

Ma, Y. (2020). Prediction of Default Probability of Credit-Card Bills. *Open Journal of Business and Management*, *08*(01), 231–244. https://doi.org/10.4236/ojbm.2020.81014

Madane, N., & Siddharth, N. (2019). Loan Prediction Analysis Using Decision Tree. *Journal of the Gujarat Research Society*, *21*(14), 214–221.

Motwani, A., Bajaj, G., & Mohane, S. (2018). Predictive Modelling for Credit Risk Detection using Ensemble Method. *International Journal of Computer Sciences and Engineering*, *6*(6), 863–867. https://doi.org/10.26438/ijcse/v6i6.863867

Mustapha, M. (2019). Food Insecurity and Coping Strategies among Rural Households in Niger State, Nigeria. *Lapai Journal of Economics*, *3*(1), 92–107.

Mythily, R., & Banu, A. W. (2017). Feature Selection for Optimization Algorithms: Literature Survey. *Journal of Engineering and Applied Sciences*, *12*(1), 5735-5739.https://doi.org/10.36478/jeasci.2017.5735.5739

Nalić, J., & Švraka, A. (2018). Using data mining approaches to build credit scoring model: Case study - Implementation of credit scoring model in microfinance institution. *2018 17th International Symposium on INFOTEH-JAHORINA*, East Saravejo, 1–5. https://doi.org/10.1109/INFOTEH.2018.8345543

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*(DEC). https://doi.org/10.3389/fnbot.2013.00021

Ndagi, A. H., Kolo, I. N., Yabagi, A. A., & Garba, Y. (2016). Adoption of Production Technologies By Lowland Rice Farmers in. *International Journal of Agricultural Extension*, *04*(01), 49–56.

Olanipekun, I. O., Olasehinde-williams, G. O., & Alao, R. O. (2019). Science of the Total Environment Agriculture and environmental degradation in Africa : The role of income. *Science of the Total Environment*, *692*, 60–67. https://doi.org/10.1016/j.scitotenv.2019.07.129

Patel, B., Patil, H., Hembram, J., & Jaswal, S. (2020). Loan default forecasting using data mining. *2020 International Conference for Emerging Technology, INCET 2020*, Pakistan, 7–10. https://doi.org/10.1109/INCET49848.2020.9154100

Pedregosa, F., Michel, V., Grisel O., Blondel, M., Prettenhofer, P., Weiss, R., Duchesnay Es. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research*, 12(1). Retrieved from http://scikit-learn.sourceforge.net.

Semiu, A., & Gilal, A. A. R. (2019). A boosted decision tree model for predicting loan default in P2P lending communities. *International Journal of Engineering and Advanced Technology*, *9*(1), 1257–1261. https://doi.org/10.35940/ijeat.A9626.109119

Sorzano, C. O. S., Vargas, J., & Montano, A. P. (2014). *A survey of dimensionality reduction techniques*. *Arxiv,*1–35.

Sulaimon, M. (2021). Agricultural credit guarantee scheme fund (ACGSF) and agricultural performance in Nigeria: A threshold regression analysis.*Journal of Applied Statistics*. Retrieved from https://mpra.ub.uni-muenchen.de/id/eprint/ 105564.

Velliangiri, S., Alagumuthukrishnan, S., & Thankumar Joseph, S. I. (2019). A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science*, *165*, 104–111. https://doi.org/10.1016/j.procs.2020.01.079

Xie, H., Li, J., & Xue, H. (2017). A survey of dimensionality reduction techniques based on random projection. *ArXiv*, 1–10.

Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and Its Applications*, *534*, 122370. https://doi.org/10.1016/j.physa.2019.122370

Zoran, E. (2019). Predicting Default Loans Using Machine Learning (OptiML). *27th Telecommunications Forum TELFOR*, IEEE Xplore.*7*, 1–27.