Hawassa University Institute of Technology

Faculty of Informatics

Department of Computer Science

MSc Thesis

Maize Crop Yield Prediction Using Machine Learning Techniques

By

Kedija Abdurhman

Advisor: Efrem Yohannes (Ph.D)

A Thesis Submitted to the Faculty of Informatics, Department of Computer Science Hawassa University Institute of Technology in Partial Fulfillment for the Requirement of Master of Science Degree in Computer Science.

Hawassa, Ethiopia

November 29, 2023

APPROVAL SHEET-I

This is to certify that the thesis entitled, “Maize Crop Yield Prediction Using Machine Learning Techniques” submitted in partial fulfillment of the requirements for the degree of Master's with specialization in Computer Science, the Graduate Program of the Faculty of Informatics, and has been carried out by Kedija Abdurhman. Therefore we recommend that the student has fulfilled the requirements and hence hereby can submit the thesis to the department.

_____	_____	_____
Name of major advisor	Signature	Date
_____	_____	_____
Name of co-advisor	Signature	Date

SCHOOL OF GRADUATE STUDIES
HAWASSA UNIVERSITY EXAMINERS' APPROVAL SHEET
(Submission Sheet-2)

We, the undersigned, members of the Board of Examiners of the final open defense by Kedija Abdurhman have read and evaluated his/her thesis entitled "Maize Crop Yield Prediction Using Machine Learning Techniques " , and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirements for the degree.

Name of Major Advisor	Signature	Date
Name of Internal Examiner-I	Signature	Date
Name of Internal Examiner-II	Signature	Date
Name of External examiner	Signature	Date
SGS Approval	Signature	Date

Final approval and acceptance of the thesis is contingent upon the submission of the final copy of the thesis to the School of Graduate Studies (SGS) through the Department/School Graduate Committee (DGC/SGC) of the candidate's department.

Stamp of SGS Date: _____

Acknowledgments

I express my deep sense of gratitude to my principal advisor Dr. Efrem Yohannes and my co-advisor Assistant Professor Abdella Kemal for their constructive and valuable supports, suggestions and comments, constant encouragement and guidance for the completion of my research. I am especially thankful to my principal advisor Dr. Efrem Yohannes for his overall supports. With my immense pleasure and profound sense, I express my heartfelt gratitude to Mr.Teshale Yohannes the Vice Head of Hadiya Zone Agricultural Department and Mr. Muhamed Faiso the expert of Hadiya Zone Agricultural Department for their full commitment to give me the data of the last 20 years. Also I would like to thank the Ethiopian metrological institute SNNPRS center for giving me metrological data for my research. I would like to thank my husband Nuriye Sule for his moral, financial and material support.

List of Abbreviations

ANN	Artificial Neural network
DT	Decision Tree
GOE	Government of Ethiopia
GDP	Gross Domestic Product
ML	Machine Learning
MAE	Mean Absolute Error
MSE	Mean Squared Error
MA	Ministry of Agriculture
PCA	Principal Component Analysis
R^2	Coefficient of Determination
RMSE	Root Mean Squared Error
SVM	Support Vector Machine

Table of Contents

List of Figures

List of Tables

Abstract

Maize is one of the main crops cultivated all throughout the world, including in Ethiopia. However, the production of maize changes widely based on many factors, such as weather, soil quality, and fertilizer usage. Predicting maize yields is crucial for farmers because it allows them to make informed crop management decisions. Machine learning approaches have shown promise in predicting crop productivity in recent years. The goal of this thesis is to explore the utilization of ensemble methods, namely Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Decision Trees (DT), in the context of maize yield prediction. Ensemble methods involve combining predictions from multiple models to enhance accuracy and fortify the reliability of the forecasting process. The dataset used in this study was compiled from data collected between 2003 and 2022 on several aspects such as weather, soil quality, and maize production. Before developing the ensemble techniques, the dataset was preprocessed and the features were normalized. The results of this research show that ensemble techniques can potentially be employed for predicting maize yields with great performance. The MAE was 0. 0025, the MSE was 0. 0032, the RMSE was 0. 0057, and the R2 was 0. 9928. The results show that meteorological factors like rainfall and temperature have a considerable impact on maize yields. Soil quality was also recognized as an important factor influencing maize crop production by the model. The research demonstrates that ensemble techniques could potentially be used to accurately predict maize yields. Farmers can use the study's findings to informed decisions about their agricultural practices. The research also emphasizes the significance of meteorological conditions and soil quality in predicting maize yields.

Keywords: Agriculture, Machine learning, Ensemble model, Maize crop, Yield prediction.

Chapter One

Introduction

1.1. Background

Ethiopia is one of the countries where the economy primarily depends on agriculture. Cattle and agricultural products are the backbones of Ethiopia's economy [1]. Since most Ethiopians live in rural regions and depend on agriculture for their primary source of income from raising livestock and crops, agriculture plays a role in the country's rural development [1]. 15–17% of the GOE's expenses are directly related to agriculture, which also produces 47.8% of the nation's GDP, more than 80% of export value, and 85% of the population's means of subsistence [2], [3]. A significant section of Ethiopia's rural population depends on maize as a major crop for both food and money. However, a lot of variables affect maize production, such as weather, fertilizer application, seed quality, soil composition, and climate. Because of these factors, farmers who wish to increase crop productivity and revenue must be able to accurately anticipate maize yields in order to plan and manage agricultural resources.

The ensemble techniques combine various machine learning algorithms that can be applied to a wide range of problems. One practical application of technology is the prediction of crop output based on a variety of input variables. The Hadiya Zone of the SNNPR in Ethiopia serves as a source of input data for this study's prediction of maize output. The model has a high degree of accuracy in predicting maize production. By giving farmers precise yield forecasts, the algorithm may potentially be utilized to increase Ethiopia's maize production. As it can precisely capture the difficult interactions between numerous parameters [4] and [5], it is an effective method for predicting maize yields.

By combining the output from multiple models to get better results, the performance of yield prediction can be greatly improved by using ensemble approaches. The ensemble techniques in this study include Decision Trees (DT), Artificial Neural Networks (ANN), and Support Vector Machines (SVM). When creating and executing artificial systems, an artificial neural network (ANN) is a network based on mathematical computations that attempts to resemble the behavior of networks observed in neurons of the central nervous systems of either humans or animals [6].

(7). The goal of this study is to see whether this method can be used to anticipate the production of maize crops, which has recently showed a great deal of promise for predicting agricultural productivity. Training datasets are used to train the various models. Because they combine the output of multiple predictor models to enhance prediction performance, certain predictors are classified as ensemble predictors among all predictors [8].

The approach is interesting in that it develops a model for prediction that helps the farmer increase maize yield. The primary goal of the thesis is on techniques, problems, and approaches for enhancing agricultural productivity, as well as how various data may be used to build a prediction model that can use an ensemble technique to boost yields from agriculture. The data set is also prepared using the thesis, and related literature is reviewed. In order to estimate the yield of maize, multiple machine learning models are developed in this thesis employing several types of variables, including weather, soil, and crop yield history. Additionally, it examines and evaluates earlier research that has already been done. An ensemble models are a group of fundamental learners. Each learner tries to learn the desired function. Their outputs are combined using an aggregation rule to get a final prediction [9].

A distinct collection of testing data is used to verify the accuracy of these models after they have been trained using a set of training data. The past data on weather patterns, soil characteristics, and maize productivity is used to train the ensemble models. The thesis evaluates the effectiveness of decision trees, support vector machines, and artificial neural networks. The results of the thesis provide useful information on the factors influencing Hadiya Zone maize output and aid farmers in making decisions about crop management and resource allocation.

1.2. Statement of the Problem

There are many challenges to increasing food security and supplying adequate and safe food supplies [3]. Among these problems are drought, conventional farming, inadequate technology use, and a lack of a platform that helps with production prediction. It is difficult to increase crop productivity when there is a weak agricultural and marketing infrastructure [2],[9]. On the other hand, Ethiopians have difficulty in using modern agricultural techniques and technology. Therefore, it is necessary to examine the farming methods used in the country. The researcher

needs to examine the unresolved technological issues in agricultural sector to increase crop productivity.

It is preferable to study Ethiopian agricultural practices and develop a yield prediction model. To advocate for better agricultural practices, the thesis must study the traditional agricultural practices used in the Hadiya zone. the ensemble techniques has utilized to predict agricultural yields, help farmers, plan for food storage and distribution, and help policymakers focus on the most vulnerable communities in order to ensure food security.

The main problems with the earlier studies, however, were the varying dataset sizes, parameter types, and parameter counts, as well as the methods used to assess the algorithms' effectiveness [10], [11]. Identifying the factors that influence agricultural productivity is another issue. Since the features extracted from data to create conventional machine learning models cannot be optimized, the performance for yield prediction may not be as good as in the majority of previous publications. Similar to other regions of Ethiopia, the traditional method, this is frequently time-consuming and requires considerable data collection and analysis, is the crop production prediction technique currently in use in the Hadiya zone Even if completed, its accuracy is in doubt. Due to the connection of multiple factors and the inherent uncertainty of agricultural systems, it is difficult to predict maize yields in Hadiya Zone. Traditional techniques often perform poorly because of inadequate performance due to their inability to precisely capture the complex relationships between input variables and output results.

1.3. Research Question

The thesis attempt to answer the following research questions:

RQ1: What are the most important factors that affect maize yield prediction?

RQ2: What are the data pre-processing techniques for improving the performance of maize yield prediction using machine learning?

RQ3: What are the most effective machine learning models for predicting maize yield?

1.4. Objectives

1.4.1. General Objective

The overall objective of this thesis is to design and develop ensemble model to accurately predict maize yield using data from Hadiya Zone in SNNPR, Ethiopia.

1.4.2. Specific Objectives

To meet the general objective, the thesis targets the following specific objectives:

- To identify the most important factors that affect maize yield prediction.
- To identify data pre-processing techniques for improving the performance of maize yield prediction using machine learning.
- To develop and test the most effective ensemble model for predicting maize yield.
- To compare the performance of the predictive model with the existing model.

1.5. Scope of the Study

Geographically, the Hadiya zone in the SNNPRS, Ethiopia, is the only place we restrict the scope of the data collection. They restrict the thesis's subject matter to machine learning-based yield prediction for maize crops. The creation of a predictive model that forecasts maize production using a machine learning strategy is the sole topic of the thesis. We restrict the model to the datasets used for testing and training, so it might not be able to predict yields for other datasets with any degree of performance.

The study focuses on the application of machine learning methods to forecast maize crop yield using data from Ethiopia's Hadiya Zone. The study trains and tests machine learning models using data on crop yields in the past as well as soil, climate, and other environmental parameters. The study assesses the effectiveness of various machine learning models and contrasts them with conventional approaches to agricultural yield prediction.

1.6. Significance of the Study

This thesis seeks to investigate the possibilities of crop yield prediction using machine learning algorithms. One of Ethiopia's most important crops, maize productions, is crucial for both food security and economic growth. However, a number of variables, such as weather, soil quality, and management techniques, greatly affect the yield of maize. Crop production estimates made by Machine learning models have the potential to be accurate, which can be used to guide decisions and raise agricultural productivity. This thesis examines the availability of using Hadiya Zone data to estimate maize crop yield using machine learning algorithms. It evaluates the data available for maize crop production in Ethiopia and analyzes the literature that has already been written on machine learning models' uses in crop yield prediction. The theses conclude by discussing the ramifications of the results and outlining suggested directions for additional study.

The theory is also employed to supply wholesome food and advance the nation's social and economic advancement. By reducing the amount of maize purchased and raising the yield of maize production, it is used to increase output for the advancement of the nation's economy and to devalue the currency. In general, the farmer, the government, and the researchers gain from the thesis. The goal of the study is to advance the development of precise and effective methods for forecasting maize production. The findings can potentially be used as a springboard for additional investigation into the application of machine learning methods for predicting maize production in different environments and crops. The importance of the maize crop in Ethiopia and the demand for accurate yield projections serve as its driving forces.

The researcher developed accurate and reliable prediction models for the prediction of maize yield using these models. It employs maize yield prediction to assist farmers in making knowledgeable decisions regarding crop management techniques and to assist policymakers in developing workable plans to guarantee food security in the region. The technique that this thesis seeks to design will raise the nation's GDP while simultaneously raising agricultural production. For efficient agricultural planning and management, crop production must be predicted accurately. It can support farmers in making knowledgeable decisions on crop choice, planting season, and fertilizer use, as well as policymakers in creating efficient agricultural policies.

Traditional approaches to agricultural production prediction, however, can take a long period and involve substantial data gathering and processing. A variable alternative is provided by machine learning techniques, which can evaluate vast volumes of data and spot complex patterns that may not be visible using more conventional techniques.

1.7. Limitation

The main issue with the thesis was obtaining the data, which led to several issues. Even when a researcher is certain they can solve a barrier, they are still subject to the following limitations: money limitations, time limitations, the involvement of relevant organizations during data collecting, reference materials, and internet access.

The production of maize is influenced by a wide range of factors, including soil, weather patterns, the prevalence of diseases, and agricultural management strategies. Machine learning models may find it challenging to accurately describe these factors' interactions and nonlinear linkages. If specific features or linkages are overlooked or oversimplified, the predictive value of the models may be diminished.

1.8. Thesis organization

Following are the concepts that make up the thesis's organization: The literature and relevant work for the subject are discussed in Chapter 2 of the book. It talks about crop production, machine learning, and Ethiopian farming. The third chapter is dedicated to the study's approach. It goes over the preparation of the dataset, its description, how features are chosen, how algorithms are found, and how metrics are assessed. A discussion of the experimental analysis is covered in the fourth chapter. The experimental set-up, model development, testing, training, assessment, maize yield prediction, and algorithm comparison are all covered in this chapter. The conclusion and advice of the article are covered in Chapter 5.

Chapter Two

Literature Review

2.1. Introduction

In this chapter, we'll look at how crucial a role maize production plays in preserving our way of life and access to food. This chapter provides a thorough assessment of the literature on the prediction of maize production with an emphasis on the use of ensemble approaches like support vector machines (SVM), artificial neural networks (ANN), and decision trees (DT). The assessments will outline the overall framework for the inquiry, explain the existing level of knowledge, and point out any gaps in the body of literature.

2.2. Agriculture in Ethiopia

In Ethiopia, agricultural productions are mostly dependent on rain-fed, about 95% of the total farming land is dependent on rain[12]. Agriculture is commonly practiced in livestock and summer season crop productions. Agriculture in Ethiopia is used for livestock and farming crops. But this thesis focuses on crop production. In Ethiopia, the farmers produce crops in different ways like irrigation, Maher, and Belg. But mostly use Maher agricultural production techniques.

Commercial agriculture in Ethiopia began during the imperial era. All of the property was given to the public in 1974, and numerous state farms were established [13]. Different elements like seasonal change, temperature, humidity, sunshine and rainfall, soil type, and fertilizers have an impact on agriculture in Ethiopia. In Ethiopia, urea and dap are the two most used fertilizer types. A variety of factors, including land and crop production, crop type choice, religion and cultural practices, capital level of investment, irrigation technique, and tillage operations, all have an impact on crop productivity in Ethiopia [12],[14]. The current approaches make use of intercropping, irrigation facilities, crop rotation, chemical fertilizers, crop rotation, crop seed improvement, irrigation facilities, and inter-cropping[15]. The adjustment of fertilizer usage is the main thing that improves crop production and Prevents soil salinity and sandy.

2.3. Crop Production

Ethiopia is one of the largest agricultural countries in Africa, with a diverse range of crops grown across different agro ecological zones [13]. Crop production is a major source of livelihood for the majority of the population and contributes significantly to the country's economy[12],[14]. However, the sector faces several challenges, including low productivity, poor infrastructure, and climate variability. In recent years, there has been a growing interest in improving crop production in Ethiopia, through various interventions such as improved seeds, fertilizers, and irrigation systems. Several studies have investigated the factors affecting crop production in Ethiopia and identified the key drivers of yield variability.

In addition to these elements, there has been an increase in interest in using contemporary technology for crop production in Ethiopia, such as remote sensing and machine learning. More than 70% of the population in Ethiopia is employed in crop production, which contributes significantly to the country's gross domestic product (GDP) [12]. Cereals, legumes, oilseeds, fruits, and vegetables may all be produced in Ethiopia thanks to its broad agricultural industry.

The Crop prediction is being processed by loading the crop data sets and the metrological dataset [16], [17]. Crop prediction is used to predict the future crop yield based on different parameters like seasonal variation, location, temperature, rainfall, crop yield detail, and soil type. Crop yield prediction is used to predict the yield of crops by using different parameters like temperature, soil type, seasonal variation, and rainfall. Crop production is an activity affected by different factors like soil fertility, climate, temperature, season, and rainfall. Crop production is predicted by machine learning approaches like ensemble learning.

2.3.1. Maize Production

2020 is projected to see a production of 7.5 million tons of maize, which is Ethiopia's second-largest grain crop after teff [18]. Growing maize is a significant agricultural activity in Hadiya Zone that has a significant impact on the local economy and food security. It is well recognized for having excellent soil and a climate that are perfect for growing maize in Ethiopia's Southern Nations, Nationalities, and Peoples' Region (SNNPR). Maize is one of the most significant crops grown in the Hadiya Zone, and it is usually planted between June and September when it is rainy.

Local varieties of maize that are well suited to the environment in the area are the most frequently grown types there. Traditional farming practices are used by farmers in Hadiya Zone to grow maize. Using manual tools and plows pulled by animals are traditional approaches. Ethiopia's production of maize is a significant agricultural endeavor that benefits the nation's food security by giving local farmers access to food and revenue. In Ethiopia, maize is a significant crop for both commercial and subsistence farming [18], [19]. It is widely grown all over the nation, with the central, southern, and western regions having the largest production areas. Ethiopia's maize industry suffers a number of difficulties, including unpredictable rainfall, degraded soil and pests.

2.4. Machine Learning

Computers can use machine learning, a branch of computer science, to learn from their prior experiences and provide results based on those experiences [6],[20]. These characteristics are examined by utilizing a machine learning approach, which is then utilized to predict the yield of crop. There are various elements for the suggestion as well as the prediction of crops. An application of AI known as machine learning gives systems the capacity to learn and advance automatically from experience without being explicitly programmed [7]. Machine learning can be classified as supervised learning, unsupervised learning, reinforcement, and semi-supervised learning[21]. Supervised learning involves training a model on a labeled dataset, where the input data is paired with corresponding output labels. The goal of supervised learning is to learn a mapping between the input and output variables so that the model can make accurate predictions on new, unseen data. Unsupervised learning, on the other hand, involves training a model on an unlabeled dataset, where the input data is not paired with any output labels. The goal of unsupervised learning is to discover patterns or structures in the data, such as clusters or latent variables. Reinforcement learning involves training a model to make decisions in an environment, based on feedback in the form of rewards or penalties. The goal of reinforcement learning is to learn a policy that maximizes the cumulative reward over time.

Numerous industries, including finance, healthcare, and agriculture, use machine learning in a variety of ways, including image and audio recognition, natural language processing, recommendation systems, and predictive modeling. But in this thesis, a crop yield prediction model was created using machine learning.

2.4.1. Artificial Neural Network

In the design and implementation of artificial systems, Artificial Neural Networks (ANN) are networks based on mathematical calculations that aim to mimic the working principles of the networks that are seen in the nerve cells of the significant nervous system of an animal or human[7], [22]. Feed-forward artificial neural networks and recurrent artificial neural networks are the two different forms of artificial neural networks. A feed-forward artificial neural network is known as such, and as such, it has just one requirement: signal flow from input to output can only occur in one direction. Back loops are not permitted, though [7], [23]. It is machine learning algorithm which is learned from the training set of data or from the example. The nodes that make up a neuron, which sends signals, are many. Data is first received by the input layer of an ANN, which processes it through one or more hidden layers before sending the results to the output layer. Each neuron in the hidden layers processes the incoming data mathematically and sends the results to the subsequent layer. The weights and biases of the neurons are altered over the course of the training phase using an optimization technique like backpropagation to lessen the error between the expected output and the actual output. ANN can be applied to a variety of tasks, such as audio and picture recognition and natural language processing, and Predictive modeling is used in many industries, such as finance, medicine, and agriculture. Modeling intricate nonlinear interactions between input and output variables makes good use of it. It has shown good results in predicting crop yields, weather patterns, and other agricultural variables, and has the potential to boost agricultural productivity and food security.

Artificial neural networks (ANN) are a potential method for predicting the productivity of maize crops due to their capacity to model complex nonlinear interactions between input and output factors. It's a specific type of machine-learning algorithm that draws inspiration from the structure and function of the human brain. They are composed of interconnected nodes, or neurons, arranged in layer-organized networks. Neurons process and transmit information. An artificial neural network (ANN) can be trained utilizing historical data, such as meteorological data, soil parameters, and yield history to determine the association between these input variables and the associated maize yield.

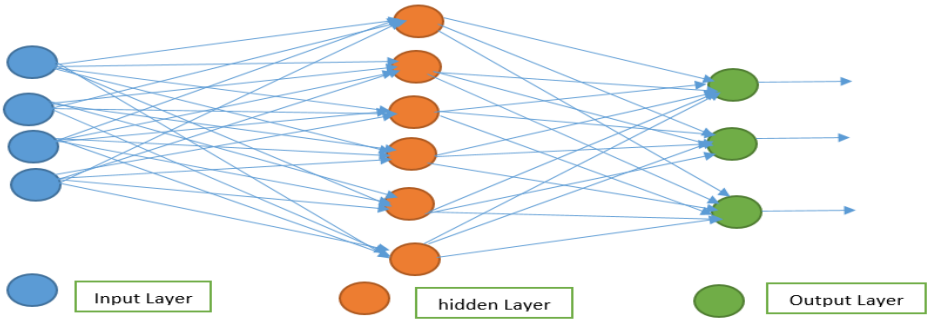

Figure 2. 1: Fully connected artificial neural network

The hidden layer has weight and threshold that are used to enhance the features of the data by performing two operations: multiplying the weight and the features of the data and adding all the results; and using a sigmoid function to generate the output. The input layer provides the hidden layer with the data in vector format. The output layer verifies the results' accuracy and provides the hidden layer with feedback. By examining a non-linear statistical data model with the interplay of the input and output parameters, ANN is utilized to process the data as the brain neuron for the construction of an algorithm that can be used to estimate crop yield.

Forward and backward propagation are the two techniques used in artificial neural network training. The training mechanism's forward propagation is its unidirectional flow. The training is output-directed and does not require feedback. The weight of the concealed layer is checked using the error function. The output is predicted using the hidden layer's activation function, whose values range from 1 to -1. Utilizing the Relu, Tanh, and ELU activation functions for the diversity of regression and predicting agricultural yield while incorporating the hidden layer, these nodes are input nodes, output nodes, and hidden layer nodes for the purpose of determining the ANN used to describe them. Fully connected feed forward neural network architecture is used in this thesis.

2.4.2. Support Vector Machine

A support vector machine (SVM) is a supervised machine learning algorithm that is used for the classification and regression of an outlier detection problem as support vector classification (SVC) and support vector regression (SVR) functions[7],[23] by considering the N-dimensional hyperplane. It is used for the yield regression and analysis for the recommendation of the crop to the farmer to increase farmer suicide. SVM develops a hyper plane and high dimensional space, which can be utilized for characterization, relapse, or different tasks[16],[24]. It is used to

analyze the dataset for regression analysis and the prediction of crops. SVM is used in many fields or areas, but is commonly used for regression and classification problems. Many of the components used for SVM classification and regression are support vectors, hyper plane, margins, and classifiers. Support vector: the data points that near hyperplane; this is a split file used to separate data in to two layers. The hyperplane is the decision plane used to divide the two objects, and it allows separating the data into two categories: The margin is the distance between two lines on the closest dataset in the different classes. SVM classifiers are used to convert the input space into the feature space.

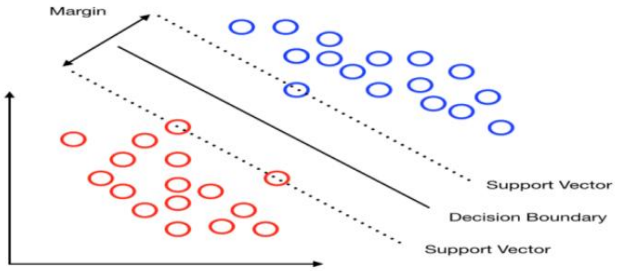

Figure 2. 2: Support vector machine

The SVM is used to predict the yield using different hyperparameters including C and gamma. Kernel Support vector machines are used for both classification and regression issues. Support vector regression can be applied in regression problems using both a linear and non-linear kernels; in the linear kernel, the data is divided into two boundaries by a straight line, whereas in the non-linear kernel, the data is not. Regression can be performed using polynomial, linear, or radial basis functions as different kernels.

Support vector regression is characterized by the use of the kernels, levels, coefficient, support vector, and margins. SVR is used to analyze the linear relationship between two or more continuous variables. The model part uses kernel functions to the control overfitting and under fitting. A kernel is a set of mathematical functions that can be used to take data as the input and give the output data by using parameters such as C (the function of regularization) and r (Gama). If the values of C and r increase then the model is over fitted, and if the values of r and C decrease, then the model is under fitted. The kernel is used to convert the data to the required format. C is used to normalize and maintain the regularization of the parameter. In support vector

machines, there are different kernel functions: polynomial, linear, nonlinear, radial basis function (RBF), and sigmoid, but we used linear kernel for this thesis.

2.4.3. Decision Tree

Two common uses for decision tree approaches are creating classification systems based on several variables and creating prediction algorithms for target variables[25]. In a decision tree, each inner node represents an attribute, while each leaf node represents an outcome. Top-level nodes have the ability to split based on the value of the attribute. The method used to split the tree is called recursive splitting. This algorithm supports decision-making. The decision tree field values are used to separate subsets of records in the database. Appling this strategy Recursively to each subset will result in all instances of each node being assigned to a single class. Decision trees are easy to understand and explain and can handle both categorical and numerical data. Overfitting, however, overfitting can occur if the tree is too complex. The decision tree method can be used for both classification and regression applications. The data is recursively split based on feature values until the resulting subsets are as pure as is practical.

The decision tree algorithm is a popular and simple supervised machine learning method for classification and regression applications[25]. This is a tree-based model that, to generate predictions, develops a hierarchy of if-else conditions based on the characteristics of the input data. Recursively dividing the input data into subsets based on feature values is how the decision tree algorithm work. The best data separation feature based on specific criteria, such as information gain or Gini impurity, is selected at each step. Until a stopping requirement is met, such as reaching a maximum depth or a minimum number of samples in a leaf node, the process continues. Divide and conquer tactics are used by decision tree learning to find the best split points within a tree by undertaking a greedy search. Recursive partitioning is another name for this method. Recursively and greedily growing the decision tree, the procedure begins by generating a single node (the root). All potential conditions are assessed and graded at each node. After that, the dividing procedure is repeated in a top-down, recursive fashion until most or all of the records have been assigned to particular class labels.

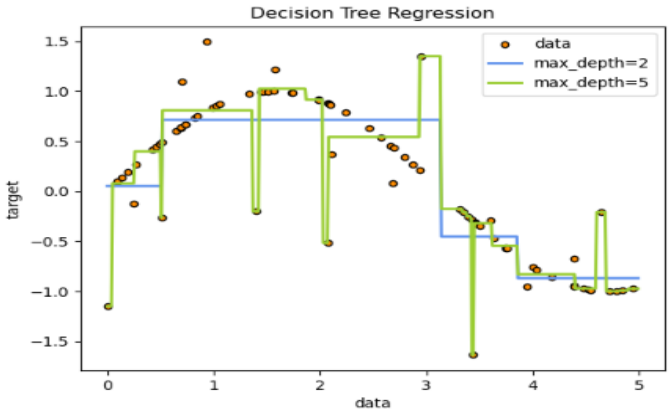

Figure 2. 3: Decision Tree

2.4.4. Ensemble Techniques

Rather than using only one model, ensemble techniques aim to increase the accuracy of outcomes in models [26]. The integrated models have greatly improved the results' accuracy. The popularity of ensemble techniques in machine learning has increased as a result. Sequential aggregation techniques and parallel aggregation techniques are the two primary categories of aggregation procedures. The base of learners is created using sequential ensemble approaches, such as adaptive boosting (AdaBoost). The foundation of subsequent learning generations encourages interdependence between individuals. The performance of the model is then enhanced by assessing learners who were previously underrepresented.

There are various types of ensemble approaches, including stacking, bagging, and boosting [26] [27]. Boosting is an ensemble technique that uses past prediction mistakes to improve future forecasts. By integrating several weak base learners into one strong learner, this technique dramatically improves the model's predictive capacity. In order to reduce training errors, the ensemble learning technique known as "boosting" turns a group of weak learners into a strong learner. A random sample of data is chosen, fitted with a model, and then trained successively in boosting, where each model tries to make up for the shortcomings of the one that came before it. Increasing efforts to create a powerful classifier by increasing the number of ineffective classifiers first, a model is created using the training set of data. The second model is then constructed in an effort to fix the flaws in the first model.

The ensemble learning technique known as bagging, often referred to as bootstrap aggregation, is frequently used to lessen variance within a noisy dataset. In bagging, a training set's data is randomly sampled and replaced, allowing for multiple selections of the same data points. These weak models are then trained independently using many data samples, and depending on the task for instance, classification or regression the average or majority of those predictions results in a more accurate estimate. Using stacking, we can train numerous models to tackle related issues and then create a new, more effective model based on the combined results. One of the widely used methods for ensemble modeling in machine learning is stacking. In order to combine different weak learners with Meta learners and make better predictions for the future, they are ensemble in simultaneously. Layered generalization is an alternate name for the ensemble approach stacking [27]. This method works by allowing a training algorithm to combine the predictions of several related learning algorithms.

The best methods for minimizing the variance of models and boosting prediction accuracy are ensemble methods. When multiple models are integrated to create a single forecast that is selected from among all other potential predictions from the combined models, the variance is eliminated. An ensemble of models integrates multiple models to guarantee that the final prediction is the best feasible one based on taking all predictions into account.

2.5. Related Articles

This paragraph describes the previous paper that was written and how it relates to the thesis. Studying and analyzing current articles that are connected to the topic is one of its goals. Therefore, it must assess earlier work and identify the paper's advantages and disadvantages.

K. Archana et al. [28] develop a voting-based ensemble classifier for the prediction of crop yield prediction. The nutrition and make-up of the soil are known to the researcher. The model was created utilizing many characteristics, including NPK, PH, and temperature. The researcher created a team-based approach for predicting crop yields. Voting-based ensemble is the best machine learning strategy for predicting agricultural yield; however, the researcher also used NB, RF, CHAID, and other algorithms.

Gandhi et al.[28] Predict crop productivity by using ANN with various factors, including crop detail, temperature, precipitation, and area of production. When predicting crop yield, which was tested on a 4-year data set from 27 districts, it had a 97.5% accuracy rate.

Mulatu and Tamir [29] developed the machine learning algorithm for the prediction of teff yield in Ethiopia. To train and evaluate various ML systems, the researcher employed a dataset containing many attributes, such as Ethiopia's climate and various growth parameters. With an accuracy of 98.38%, the researcher discovered that the convolutional neural network performed best in forecasting Teff yield. Images from satellites were utilized by the researcher. The researcher may provide more details regarding the soil characteristics and meteorological data that they used as input to the ML model. Additionally, the researcher might have provided additional details regarding the various models they compared. The research illustrates the potential of machine learning to make a significant contribution to the field of crop yield prediction.

Muthoni et al. [30] developed a machine-learning algorithm for the precise prediction of maize grain yield in Southern African conservation agricultural systems. In order to train and test machine learning methods including decision trees, random forests, and support vector regression and assess the significance of various parameters in forecasting maize grain yields, the researcher employed a three-year dataset from Zimbabwe. The study did not take into account the effect of climate variability on maize grain yields, a crucial factor affecting agricultural productivity. On the use of machine learning techniques to forecast maize grain yields in conservation agriculture systems, the researchers provide interesting insights. The study found that the RF model accurately forecasted maize grain yields in conservation farming systems in Southern Africa. Having an R2 score of 0.63

Croci et al. [31] developed a new method for predicting maize yield by combining data from multiple sources, such as satellite imaging, meteorological data, and soil data. Over a five-year period, they collected data from 11 sites across Italy. Their approach is reliable and can be used to forecast maize yield on a local or global level. Despite using a different model, the researcher found that the Gaussian process regression approach performed best, with an RMSE of 13.31%. To estimate maize production, the researchers employed a range of machine learning methods,

and they tested the effectiveness of each model on a test dataset. Since the study was carried out in Italy, it is unclear whether the findings would apply to other nations. The research used

Shah Hosseini et al. [32] develop machine learning algorithms for the prediction of maize yield and nitrate loss for sustainable agriculture and environmental protection. The dataset used to train and evaluate the machine learning algorithms is then described. It contains data on weather, soil characteristics, management techniques, maize production, and nitrate loss across a number of years. To forecast maize output and nitrate loss, the researchers use a variety of machine learning methods, including LASSO regression, Ridge regression, random forests, extreme gradient boosting, and ensemble. The study's findings demonstrate that machine learning algorithms are capable of accurately forecasting maize yield and nitrate loss, with XGBoost showing the greatest performance with 13.5% RRMSE. A significant addition to the field of sustainable agriculture is made by this publication.

Baio et al. [33] develop a machine-learning algorithm for the prediction of maize yield with spectral variables and irrigation management variables. The researchers use a dataset of historical maize yield, weather, and soil data and develop different machine learning models, including Random Forest and Gradient Boosting Artificial Neural Network (ANN), M5P Decision Tree (J48), REPTree Decision Tree (REPT), Random Forest (RF), and Support Vector Machine (SVM), to predict maize yield, and multiple linear regression (LR) was tested as a control model. The model is then evaluated on a held-out test set, and the researchers show that it can predict maize yield by Random Forest with a high degree of accuracy. The paper provides valuable insights into the use of machine learning and spectral data for crop yield prediction and has potential implications for improving agricultural practices and increasing crop productivity.

Biru and Tefera[34] look into the use of GIS and remote sensing to anticipate maize yield In the Kaffa Zone, southwest Ethiopia. The relationship between maize yield and environmental parameters like rainfall, temperature, and soil moisture was examined using satellite imagery and GIS data. Based on these variables, they create a model to forecast maize production, and they validate the model using data from the field. Using data from eMODIS_NDVI, actual and projected, the researchers found that a spectro-agro meteorological yield model with a r2 of 0.89, an RMSE of 1.54 qha-1, and a 16.7% coefficient of variation could accurately estimate maize

yield at the zone level. The authors come to the conclusion that maize yield projections based on GIS and remote sensing enhanced the data.

Reddy and Kumar [36] The study is focused on examining various aspects based on data availability using machine learning methods for crop yield prediction (CYP). Instead of using additional features, the selection of features tries to determine the best-performing features by taking into various factors including geological position, scale, and crop characteristics. Neural networks, random forests, and KNN regression methods. However, there is still potential for progress in CYP, and the study emphasizes using ML approaches while considering temperature, weather, and other factors into consideration. The study hasn't consideration soil type and fertilizer usage to help farmers make better decisions. The findings of the research proposes developing a deep learning (DL) model for CYP based on the results, emphasizing the significance of feature selection, the power of machine learning (ML) algorithms, and the potential for further advancements in crop yield prediction.

Mupangwa et al. [37] look at Machine learning for the prediction of maize yield in eastern and southern Africa. Datasets on maize yield from 2012 to 2017 were used by the researcher in Zimbabwe, Zambia, and Malawi. The researcher utilizes a variety of machine learning techniques, including MLR, ANN, and SVM. ANN performs best, with an average accuracy of 92%. The researcher suggested more research be done to look into how ML may be used to predict yields for other crops in other parts of the world. The researcher has both advantages and disadvantages. The researcher outlines the machine learning model in detail when we start with the strengths. The flaw is that it could be strengthened by providing more information about the data that was provided.

Liu et al.'s [38] study of machine learning algorithms for estimating crop yields using soil and meteorological datasets. Data on maize yield is used by the researcher to train and test a variety of machine learning algorithms, such as XGBoost, RF, and LSTM. The study makes use of crop yield data that was gathered in the US from 2010 to 2015. On climatic factors and soil characteristics for maize in the American Corn Belt, the researcher trained and tested a number of ML models. Additionally, they contrast the effectiveness of the ML models with that of the Decision Support System for Agro technology Transfer (DSSAT), a process-based model

(PBM). The study's findings demonstrate that ML models are highly accurate at forecasting crop yield.

Tesfaye et al. [39] investigate the use of machine learning and remote sensing to forecast Ethiopian smallholder wheat harvests. In the wheat region of Ethiopia, the researchers collected data from 165 farm sites, including agronomic data that was collected on-site and Sentinel-2 and Sentinel-1 satellite photos. Then, they used a variety of machine learning and deep learning techniques, such as stacked ensembles, deep learning, distributed random forests, gradient boosting machines, generalized linear models, and extremely randomized trees (XRT), to predict wheat yield from the satellite imagery. To estimate wheat yield, the researchers employed various machine learning and deep learning techniques. They then tested the effectiveness of the various models on a test dataset. Although the study was carried out in Ethiopia, it is unclear whether the findings would apply to other nations. Additionally, the study only used a modestly sized dataset; therefore, it is unclear how different the findings would be with a bigger dataset.

Cedric et al.'s [40]study uses machine learning techniques like DT, KNN, and logistic regression to forecast crop productivity in West Africa based on numerous meteorological and soil data points. The study's findings demonstrated that using soil and meteorological data, machine learning algorithms could successfully estimate crop yields. With a coefficient of determination (r2) of 95.3%, the decision tree model was discovered to be the most accurate model for predicting crop yields for six crops in West Africa, including rice, maize, cassava, seed cotton, yams, and bananas. The study also offers insightful information regarding the connections between agricultural yields in West African nations and soil and meteorological data. Findings from the study may aid in establishing

Nigam et al. [41] provide crop forecasts and crop suggestions to the farmer by examining the soil features. By utilizing the Nave Bayesian machine learning approach, the study increases the precision and effectiveness of the recommendation system based on parameters such as soil feature, crop type, farmer detail, yield detail, crop detail, and agricultural parameter (total rainfall). Based on the kind of soil, irrigation technique, pH level, size of the property, and prior experience, recommended crops are determined. This study identifies the criteria or factors that a farmer requires to field certain types of crops and the crops that are produced using various

fertilizer rates per hectare. The report offers insightful information on the use of machine learning methods to forecast crop yields in India. The study's conclusions may help anticipate crop yields more accurately and guide management choices in other places as well. To validate the results and examine the possibilities of other environmental elements and machine learning algorithms for agricultural production prediction, more research is nonetheless required. The studies' datasets could have been described in greater detail, and the machine learning techniques could have been subjected to a more thorough examination.

Bondre et al. [42]studied the prediction of crop yield, the classification of the soil, and ultimately recommended the fertilizer of the soil using machine learning techniques like SVM and random forest based on soil fertility, location, crop yield, and weather parameters. The accuracy of random forest was 86.35%, and that of support vector machine was 73.75%. The farmer uses the paper to demonstrate how they apply the fertilizer effectively and how they raise crop output.

Siva et al.'s [43] investigation into the potentials and constraints of soil serves as the foundation for their recommendations for the ideal fertilizer type, dosage, and timing. This study put out an ANN model that forecasts NPK nutrient levels and suggests the most effective treatment based on the predicted weather. An RMSE of 0.5 was reached by the generated model.

Gue et al. [44] studied time series analysis by utilizing historical data on wheat production, area, and weather variables, including rainfall and temperature. For predicting wheat yield, the study uses both statistical analysis and neural networks. With greater accuracy than existing models like the NARXNN and NARNN, the article replicates both spatial and temporal neural networks for the prediction of wheat yield.

Abraham et al.'s [45] predicts soybean production, harvested area, and yield by using ANN model within 50 years of the dataset's collection. The accuracy of the harvested area prediction is 0.944 in R2, 1.813 in MAE, and 3.915 in MSE; the accuracy of the yield prediction is 0.899 in R2, 0.158 in MAE, and 0.037 in MSE; and the accuracy of the production prediction is 0.968 in R2, 3.990 in MAE, and 21.755 in MSE.

Yadav et al.[46] Predict the agricultural crop yield using a machine learning strategy called a support vector machine. Within the 10-year dataset, multiple linear regression and random forest with various parameters, such as rainfall and temperature,

Medar et al. [47] develop the prediction model using machine learning methods like LSTM, RNN, and RF. And the accuracy of the prediction obtained by the researcher was 98.42%. For the forecast of crop yield, the model incorporates many attributes. These factors include the economy, the rule of law, the rate of production, the time of year, and geography. A one-year dataset is gathered by the researcher, but because the predicted performance is poor, she just chooses the lowest performance. This study aids farmers in boosting agricultural productivity.

Kumar [48] developed the supervised machine learning methods for crop yield prediction. In this study, crop yield prediction taken into account using historical data that takes into account elements like temperature, humidity, PH, rainfall, and crop name. The best crop yield model, when compared to other models, is obtained by the researcher using random forest (RF). The researcher builds a crop yield prediction model using a variety of machine learning techniques, including RF, DT, MLR, and NB, and uses a variety of parameters. He achieves 78% accuracy with the prediction model.

Table 2. 1: Summary of related work

Article	Data Features	Data set size	Algorithm	Predicted	Performance	Drawback
[28]	NPK, pH, temperature	20 crop type dataset	NB, RF, CHAID, and Voting-Based Ensemble	Yield Prediction,	94%	Different crop type, doesn't specify the Size of datasets and lowest performance
[49]	Temperature, rainfall, area, and season	100-year data set	LSTM, RNN, RF	Predict and classify crop	67.8%	Lowest performance
[50]	Marketplace, government police, production rate, season, region	One year data (Wheat and rice)	NB, KNN	Crop classification and prediction	91%	Lowest performance and less dataset
[51]	Temperature, rainfall, humidity, ph. and crop name	3101 instance with 5 parameter	RF, DT, MLR, NB	Crop yield prediction	78%	Lowest performance and less dataset
[52]	Evapotranspiration, season, temperature, yield detail	Four-year Rice data set	ANN	Predict crop yield	98.12 %	Less amount of dataset and less number of attribute
[53]	Area, temperature, rainfall, soil Ph	- Rice yield	Ensemble	Predict yield	0.86 R^2 0.066 MSE	Less amount of dataset and less number of attribute
[54]	Yield data, Weather data, Soil data and Management data	18 years corn dataset	ensemble models	Predict yield	7 to 20% RMSE	lowest performance

[55]	Soil characteristics, soil type, crop detail	100 rows of data	RT, CHAID, KNN, NB	Predict crop and suggest to the farmer	88%.	Lowest performance
[56]	Soil Type, pH value of the soil, NPK content of the soil, Porosity of the soil, rainfall, Surface, Temperature and season	5MB within 9000 rows	RF, NB, SVM	Predict crop yield	99.91	does not discuss the interpretability of the model
[57]	Depth, Texture, Ph, Soil Color, Permeability, Drainage, Water holding and Erosion	groundnut, pulses, cotton, vegetables, banana, paddy, sorghum, sugarcane, coriander	SVM, ANN, RT, NB	Crop Recommendation	high accuracy and efficiency	Undefined performance
[58]	Temperature, humidity, PH, Rainfall, crop	10-year data	NB	Crop prediction	97%	Least amount of dataset and feature
[7]	Temperature, perception, crop detail, and area	The four-year dataset from 27 district	ANN	Crop prediction	97.5%	Least amount of dataset and feature

With little data and few features, the existing research uses several machine learning techniques to estimate agricultural yield, which results in a model with poor performance[49], [51], [36]. A more comprehensive dataset with large number of parameters are used for this thesis. Most the earlier studies uses single machine learning algorithms but this thesis utilizes an ensemble machine learning methodology. Rather than using single model ensemble techniques can significantly increase the performance of yield prediction by integrating the results from various models to generate better results, Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Decision Trees (DT) are the ensemble techniques used in this study.

By utilizing machine learning approaches, the thesis creates a cutting-edge method for forecasting maize yield. Most of the currently published articles do not specify the hyper parameters that mean researchers randomly train a model. The hyper parameters are, however, chosen and optimized by the thesis. Depending on how much the cost function and training procedure reduce the test error; grid search is used to evaluate the model using a validation set. Furthermore, in our country, such journals don't do in-depth research.

Chapter Three

Research Methodology

3.1. Introduction

This chapter discusses a research methodology to prepare a dataset and techniques to achieve the objectives and answer the research question. This thesis develops a predictive model for agricultural production using a machine learning approach. The thesis uses content analysis to find out how machine learning raises agricultural productivity and sustainability. However, all of the resources, procedures, and goals of this thesis, as well as the model used, must be described and explained.

3.2. Yield prediction model work flow

This thesis uses data from Ethiopia's Hadiya Zone to predict maize yield. Ensemble techniques were used in the thesis by combining support vector machines (SVM), Artificial Neural Networks (ANN), and Decision Trees (DT). For the algorithm, the initial dataset has to be trained and handled differently. A Python script is used to implement this. The approach was then refined using data, and model was created. After creating the model and predicting the maize yield, we evaluated it using test set.

3.2.1. Research Design

The thesis architecture, as shown in figure 3.1, examines and constructs the research first, then gathers the dataset and preprocesses the data to enhance the quality of the data set and effectively gather the precise results we need. Using several feature extraction algorithms, preprocessing converts text input into actual vectors. Separate the dataset into training and test sets as well. Next, create a machine learning algorithms and forecast the yield of maize. See the following section for a more thorough explanation. The architecture for predicting maize yield using machine learning algorithms is shown in the image below.

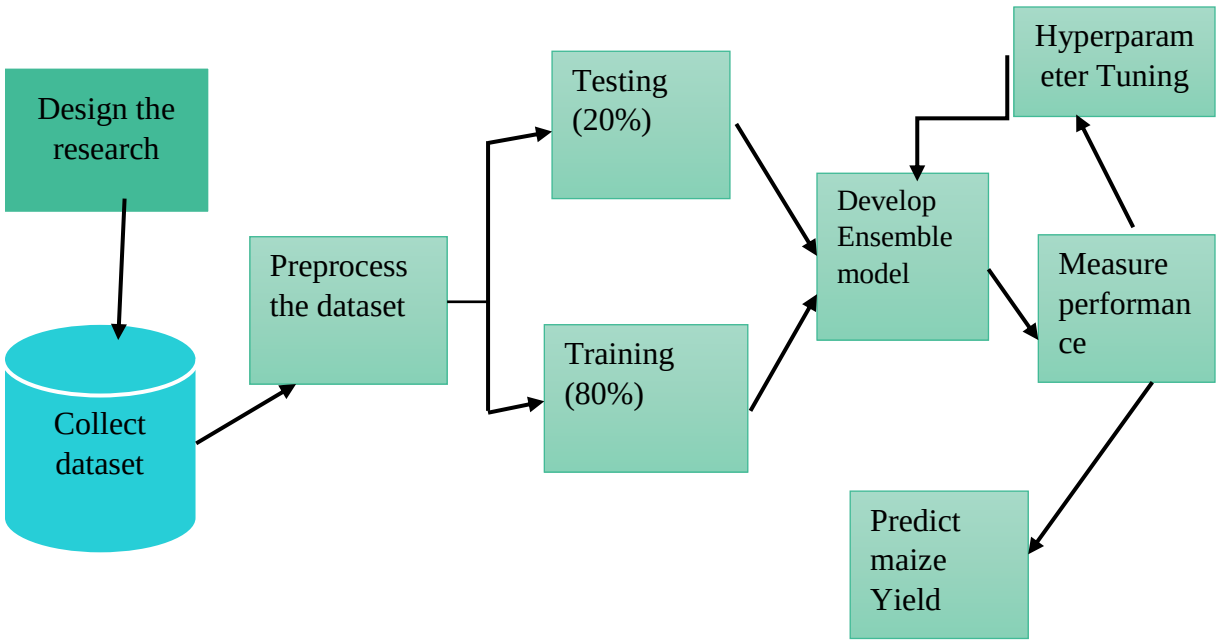

Figure 3. 1: The prediction model work flow

In this study, the researcher chooses an experimental research strategy since it allows for the incorporation of several variations and facilitates comparison analysis. Our goal is to improve prediction methods by investigating a variety of factors. It is essential to define a methodology through study design in order to construct a model. In order to do this, the issue at hand must be carefully examined, the investigation's scope must be determined, and the factors that affect the model's creation must be chosen. The study identifies the research area, focuses on the particular issue there, and chooses the best parameter for predicting the yield of maize. Furthermore, the prediction of maize yield in the context of Ethiopia is thoroughly investigated and examined in this study.

3.2.2. Types of data source

After identify the research area and the problem in the maize yield prediction we need to collect the dataset and identify the types of the dataset. The research used secondary sources of data to achieve the objective of the study. The secondary sources of data are different documents and Crop Production Reports from the Ethiopian meteorology Institute and Hadiya Zone agriculture office.

3.2.3. Sample Size

The study used 20 years of weather data, data on annual maize yields per hectare, soil type and amount of fertilizer used per hectare from Hadiya zone covering 13 woreda and the zonal agricultural department and from Ethiopian metrology institute SNNPR center. The thesis uses 4121 raw of data with 14 attributes.

Table 3. 1: Dataset Sample

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4121 entries, 0 to 4120
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Year           4121 non-null   int64
 1   Woreds         4121 non-null   object
 2   Crop           4121 non-null   object
 3   Area/hectar    4121 non-null   float64
 4   Production     4121 non-null   float64
 5   Urea/kg        4121 non-null   float64
 6   Dap/kg         4121 non-null   float64
 7   Soil           4121 non-null   int64
 8   Tmax           4121 non-null   float64
 9   Tmin           4121 non-null   float64
 10  RF             4121 non-null   float64
 11  hum            4121 non-null   float64
 12  SS             4121 non-null   float64
 13  Yield/kuntal   4121 non-null   float64
dtypes: float64(10), int64(2), object(2)
memory usage: 450.9+ KB
```

3.2.4. Data collection method

The secondary information could be gathered from Crop Production Reports from the Ethiopian Meteorological Institute SNNPRS center and from Hadiya Zone Agriculture Office, that relate to metrological information (like rainfall, temperature, sunshine, and humidity) and the crop yield history (types of crop, area per hectare, production per hectare,

and the usage of fertilizer per hectare), and the types of soil (Loam Soil, Clay Soil, and Sandy Soil).

Table 3. 2: The collected dataset

	Year	Woreds	Crop	Area/hectar	Production	Urea/kg	Dap/kg	Soil	Tmax	Tmin	RF	hum	SS	Yield/kuntal
0	2003	shashego	Maize	85.25	15.0	0.334216	0.82120	0	18.980	7.12	169.14	72.40	4.46	1278.75
1	2003	shashego	Maize	21.00	19.0	2.125000	2.75000	1	19.025	7.85	97.50	141.55	8.10	399.00
2	2003	shashego	Maize	85.25	16.0	0.334216	0.82120	2	19.025	7.85	97.50	141.55	8.10	1364.00
3	2003	shashego	Maize	121.00	17.0	2.345000	2.75000	0	19.620	7.72	141.30	70.36	6.40	2057.00
4	2003	shashego	Maize	61.00	10.0	0.334216	0.82123	1	22.400	6.83	111.50	70.00	8.40	610.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4116	2022	duna	Maize	3.00	24.0	0.334216	0.82120	1	19.620	7.72	141.30	70.36	6.40	72.00
4117	2022	duna	Maize	78.00	21.0	0.334216	0.82120	2	22.400	6.83	111.50	70.00	8.40	1638.00
4118	2022	duna	Maize	78.75	19.0	0.334216	0.82120	0	17.500	8.60	281.45	83.90	3.75	1496.25
4119	2022	duna	Maize	39.00	16.0	0.334216	0.82120	1	19.480	7.13	142.00	66.80	4.88	624.00
4120	2022	duna	Maize	65.00	28.0	0.334216	0.82120	2	18.980	7.12	172.20	72.50	3.73	1820.00

[4121 rows x 14 columns]

Table 3. 3: The attribute description

No	Attribute	Description
1	Year	The year in which the crop was grown is an important parameter as it can help identify trends and patterns in crop production over time.
2	Woreda	Woreda are Ethiopia's administrative divisions, followed by zones and regional states, and further subdivided into kebele neighborhood associations
3	Area	The area of land being used for crop production is an important parameter. The system need to consider the size of the land to make accurate recommendations
4	Production	The total amount of crop harvested and it is an important parameter as it can help identify trends and patterns in crop yields. It is measured by tone per hectar
5	Soil type	The type of soil in the area where the crops will be grown is an important factor in determining which crops will thrive.
6	Rainfall	The amount of rainfall in the region is an important factor in crop selection. The system need to consider the amount of rainfall required for the crop to grow.
7	Sunshine	The amount of sunshine in the region is an important factor in crop selection.
8	Temperature	The temperature of the region is another important factor in crop selection. The system will need to consider factors such as the minimum and maximum temperature required for the crop to grow.
9	Humidity	The humidity of the region is another important factor in crop selection.
10	Yield	The amount of crop harvested per unit of land and it is measured by kilogram/per hectar

11	Fertilizer	Urea	The amount of urea fertilizer used is an important parameter as it can impact the growth and yield of the crop.
		Dap	The amount of DAP fertilizer used is an important parameter as it can impact the growth and yield of the crop.

3.2.5. Data Preprocessing

Preprocessing is the process of analyzing and correcting the data set[59]. Preprocessing involves adding the missing values, the correct set of data, and extracting the functionality. Preprocessing is normalizing the features of data and the range of dependent and independent data[60]. It needs to replace missing values, correct spelling, and avoid redundancy for the preprocessing phase. This procedure is used to drop and delete unused columns, used to replace missing values, used to change the data types of the columns to the required format, and used to remove data redundancy. In the preprocessing step, the dataset would be identified based on their correlation. Once the data has been collected, the next step is to select relevant features. These features can include temperature, rainfall, soil type, and other environmental factors that affect crop growth. Clean and preprocess the data to ensure that it is in a suitable format for use in machine learning models[59]. This may involve data normalization, feature selection, and data augmentation techniques[61]. To improve the performance of maize yield prediction using machine learning, we are used several effective data preprocessing techniques. These techniques include: data cleaning, normalization, feature encoding and feature selection.

Handling redundant data points and missing values that could harm the model is known as data cleaning. Additionally, one can use normalization to standardize or normalize numerical features to ensure that they have the same scale. Two different types of scaling procedures are used in normalization. Normal distribution is followed by standard scaling. As a result, it adjusts the data to a unit variance and sets mean=0. This traditional strategy might be better if the data had a normal or Gaussian distribution, but it does not provide balanced feature scales when there are outliers. The second scaling methods include One of the most used scaling techniques is minmax scaling. It translates features by using the minimum and maximum values of the original data to scale data values to a range between 0 and 1. We employed the minmax scaling method for this

thesis. Due to the fact that our data does not follow a normal distribution, this strategy preserves the relative order and proximity of the data points while also reducing variance and amplifying the impact of outliers. Another technique is label encoding is one example of feature encoding, which transforms category data into numerical representations appropriate for machine learning algorithms.

The final preprocessing method is feature selection, which is essential for decreasing noise and improving model performance by locating and choosing the most important features. The selection of features is crucial for reducing the computational complexity of the model. Additionally, it improves the algorithms' capacity for prediction by focusing on the most important factors while eliminating the unnecessary and unimportant ones. For this, methods like feature importance ranking and correlation analysis are frequently employed. These preprocessing methods can greatly enhance the performance of models that forecast maize yield.

3.2.5.1. Dataset correlation

A statistical concept called correlation assesses the connection or association between two or more variables [55]. It's crucial to comprehend how changes in one variable affect changes in another. From -1 to +1, the correlation coefficient can be calculated [56]. If one measure rises, the other measure also rises linearly, this is the case when the correlation value is 1, which indicates a complete positive correlation. A correlation value of -1, on the other hand, denotes a complete negative correlation, in which one measure rises while the other decreases linearly. There is no association between the variables, and there is no evidence of a linear relationship, when the correlation value is zero [57]. The two variables change in the same direction when there is a positive correlation. In a neutral correlation, no correlation exists between the changes in the variables in a neutral correlation; no correlation exists between the changes in the variables. Variables change in the opposite way when there is a negative association. A positive correlation implies that as one variable rises, the other rises as well. A negative correlation, on the other hand, suggests that as one measure rises, the other falls.

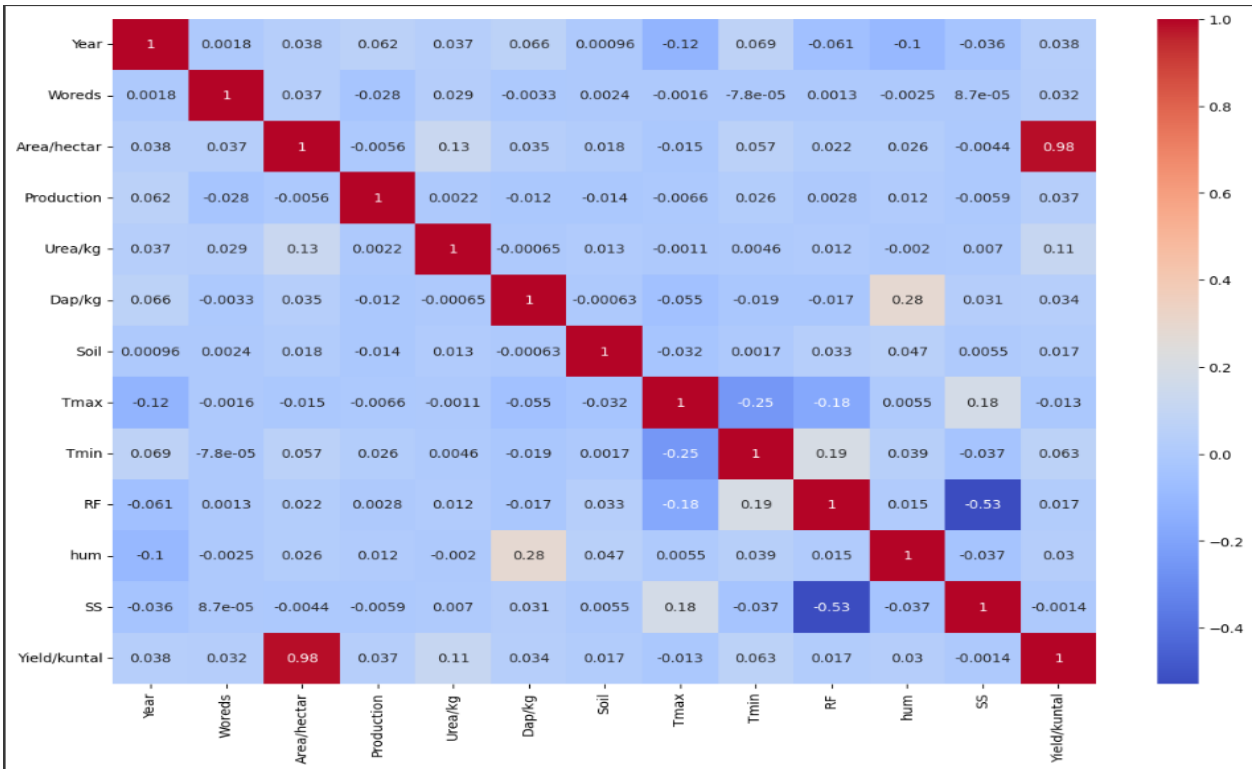

Figure 3. 2: dataset correlation

From the heat map, we can see that the features with the positive correlation to maize yield are area, humidity, temperature, soil type, dap, rainfall and urea. These features have a positive correlation with maize yield, indicating that higher values of these features are associated with higher maize yields. On the other hand, sunshine has negative correlation with maize yield.

3.2.5.2. Dataset Normalization

Normalization is the process of scaling the data so that all features are on the same scale[62]. This is important because many machine learning algorithms, including neural networks, are sensitive to the scale of the input features. If the features are on different scales, some features may dominate the others, leading to biased results. Normalization ensures that all features are on the same scale, which can improve the accuracy and performance of the machine learning model[63].

$$X_{Normalized} = \frac{(X - X_{Mimimum})}{(X_{Maximum} - X_{Mimimum})} \text{----------eq (3.1)}$$

Normalized dataset is important for the crop prediction using machine learning because it ensures that all features are on the same scale, which can improve the accuracy and performance of the machine learning model. By normalizing the dataset, we can avoid issues where some features may dominate others due to differences in scale, leading to biased results.

Table 3. 4: Normalized dataset

	Year	Woreda	Area/hectar	Production	Urea/kg	Dap/kg	Soil
0	0.0	0.666667	0.000092	0.045614	0.000023	0.201797	0.0
1	0.0	0.666667	0.000694	0.059649	0.000168	0.760870	0.5
2	0.0	0.666667	0.000092	0.049123	0.000023	0.201797	1.0
3	0.0	0.666667	0.000710	0.052632	0.000186	0.760870	0.0
4	0.0	0.666667	0.000066	0.028070	0.000023	0.201806	0.5
...	...	...	...	...	...	...	...
4116	1.0	0.166667	0.000419	0.077193	0.000023	0.201797	0.5
4117	1.0	0.166667	0.000085	0.066667	0.000023	0.201797	1.0
4118	1.0	0.166667	0.009992	0.059649	0.000023	0.201797	0.0
4119	1.0	0.166667	0.000425	0.049123	0.000023	0.201797	0.5
4120	1.0	0.166667	0.000071	0.091228	0.000023	0.201797	1.0

	Tmax	Tmin	RF	hum	SS	Yield/kuntal
0	0.429293	0.062829	0.431998	0.457086	0.315417	0.000048
1	0.434975	0.071965	0.240652	0.917166	0.830269	0.000455
2	0.434975	0.071965	0.240652	0.917166	0.830269	0.000051
3	0.510101	0.070338	0.357639	0.443513	0.589816	0.000417
4	0.861111	0.059199	0.278045	0.441118	0.872702	0.000023
...	...	...	...	...	...	...
4116	0.510101	0.070338	0.357639	0.443513	0.589816	0.000347
4117	0.861111	0.059199	0.278045	0.441118	0.872702	0.000061
4118	0.242424	0.081352	0.731971	0.533599	0.214993	0.006551
4119	0.492424	0.062954	0.359509	0.419827	0.374823	0.000235
4120	0.429293	0.062829	0.440171	0.457751	0.212164	0.000068

[4121 rows x 13 columns]

3.2.6. Split the Dataset

After preprocessing, split the data set into a training data set and a test data set based on the specified split ratio. It's crucial to separate the data into the training set and the testing set while learning dependence from data in developing predictive model in order to prevent overfitting. The data from the testing set is used to determine the model's correctness after it has been trained

on the training set. However, the choice of the split ratio depends on the size and quality of the dataset as well as the complexity of the model being developed. Striking a balance between sufficient training data and a reasonable amount of test data is crucial for evaluating the model's performance accurately. For this study we have used 80/20 split ratio, it is commonly used split ratio which means 80% used for training and 20% for testing. The reason for using an 80/20 split is that it provides a good balance between having enough data to train the model and having enough data to test the model's performance.

Data scientists are frequently instructed to develop supervised models in machine learning on big training datasets and then test them on smaller datasets. The larger the dataset used for training, the more effectively the model learns, which is why it is always selected to be larger than the test dataset. It should always yield better results for the model than the testing set as the model was trained on the training set rather than the testing set. Because of these reason we have chosen the 80/20 split ratio.

3.2.7. Apply the Machine Learning Method

When dealing with machine learning algorithms, we want to see how well our machine learning model picks up new information and adapts to it. Prediction errors, which are a measure of the machine learning model's performance and accuracy, are what we actually refer to when we discuss it. Overfitting and under fitting, which are majorly responsible for the poor performances of the machine learning algorithms. The choice of model depends on the type of data available and the specific problem being addressed. Train the machine learning model using the preprocessed data. This involves feeding the model the input data and adjusting the model's parameters to minimize the error between the model's predictions and the true crop yield data. We develop the model after preprocessing the dataset, and splitting dataset in to training and testing the dataset. Basically, we evaluate different models to address optimization problems when working on a machine learning problem with a given dataset. In machine learning, our goal is to create predictive models that forecast the result for a given an input data. We adjust the trained model further to do this. So, in order to determine which candidate model performs the best, we assess their performance.

The ensemble model employed to combines judgments from other models to enhance performance as a whole. In order to enhance model performance, three different algorithms

ANN, SVM, and Decision tree process—were merged in this study. On average, many predictions are made for each piece of data. This method uses an average from all the models to arrive at the final forecasts. The process that is used to create the ANN for the training and the prediction of the dataset is first to create different network input by using multiple futures that affect the maize production, then analyze the network result or output after examining the hidden networks and also analyze the testing model of the network. These processes use activation functions that range between 0 and 1. High-dimensional data handling, non-linear decision boundaries handled by the kernel and reliable generalization performance are all characteristics of SVM. The objective of SVM is to identify an ideal hyper plane that either predicts the target values with the greatest margin of error or maximally divides the samples from distinct classes. The hyper plane is also known as a decision boundary that divides the classes by maximizing the distance between the nearest samples of various classes. The hyperplane with the biggest margin is the one that SVM seeks for because it is seen to be the best option. The entire dataset is represented by the decision tree algorithm's root node. The approach constructs a child node for each partitioned component in order to continue recursively. The process of selecting the optimal feature and threshold for splitting is repeated until a halting condition is met.

To increase overall prediction performance and robustness, ensemble techniques integrate the predictions of various independent models, such as SVM (Support Vector Machines), ANN (Artificial Neural Networks), and DT (Decision Trees). Using these three models, here is a general explanation of how ensemble approaches work: First, The training data is used to individually train each model (SVM, ANN, and DT) using its corresponding algorithm. Then by employing several algorithms or by instructing each model to be trained on various subsets of the training data, ensemble strategies seek to produce diverse individual models. This diversity decreases the risk that all models will make the same mistakes by allowing them to better capture various parts of the data. The ensemble technique combines the predictions of the separate models after they have been trained to produce a final forecast.

3.2.8. Hyper-Parameter Tuning

Setting and managing the values of hyperparameters that necessitate trial and error is a significant difficulty when dealing with machine learning algorithms. The performance of any model depends greatly on its parameter values, such as the number of layers, the number of

neurons in each layer, the epoch and learning rate for ANN and c, the kernel type, and the kernel coefficient for SVM and max_depth, min_samples_split, and min_samples_leaf for DT . By comparing several combinations of hyper-parameters and selecting the one that produces the greatest outcomes, the configuration parameters of a machine learning model are used to internalize the model and forecast the model by learning from the data. Their values can be predicted from the dataset. A machine learning model uses internal parameters to internalize the model, whose values may be guessed from the provided dataset. This allows the model to be predicted by learning from the data. A hyper-parameter is a configuration parameter used to externalize a machine learning model, and its values cannot be inferred from the dataset. The selection of the hyperparameters affects performance of machine learning algorithm. To optimize the performance of the model, tuning of hyperparameters involves selecting the best combination of the parameters. There are three different types of hyperparameter tuning techniques, but the two that are most frequently used are grid search tuning techniques and random search tuning techniques. We have chosen to use grid search for our models because random hyperparameter setting makes it difficult to compare the performance of different algorithms; grid search works better by generating a grid of all possible hyperparameter values. In order to analyze the range of hyperparameters in a machine learning model, grid search is one of the tuning approaches employed. Values of hyperparameter, it is better approach which work by creating a grid of all possible hyperparameter values. Grid search is the tuning techniques that used in machine learning model to evaluate the range of hyperparameters.

3.2.9. Performance Measurement

The performance of a maize yield prediction model can be evaluated using several metrics, including the coefficient of determination (r2_score), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE).

A. Coefficient of determination (R2_score)

The coefficient of determination, also known as R2_score, is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables[64]. In the context of maize yield prediction using machine learning algorithm, R2_score is used to evaluate the performance of the model in predicting the yield of maize based on the input variables.

$$R2_{score} = 1 - (^{\text{sum of squared residuals}}/_{\text{total sum of squares}}) \text{ ----------eq (3.2)}$$

Where, the total sum of squares is the sum of the squared differences between the actual values and the mean of the actual values, the sum of squared residuals is the sum of the squared differences between the predicted values and the actual values. The R2_score is used to assess how well the model fits the data while predicting maize yield using machine learning algorithm. The ability of the model to explain a significant amount of the variation in the dependent variable (maize yield) based on the independent variables (such as weather information, soil quality, etc.) is indicated by a high R2 value. A low R2 score means that the model does not adequately account for the independent variables' contributions to the variance of the dependent variable.

B. Mean absolute error

MAE is used to assess how accurately the model predicts the yield of maize based on the input variables in the context of maize yield prediction using machine learning algorithm. MAE is used to evaluate the performance of the models performance[65].

$$MAE = \frac{1}{n}\sum_{k=0}^{n} \quad yk - yk \vee - - - - - - - - - eq(3.3)$$

C. Mean squared error (MSE)

The mean squared error (MSE) is used to assess how well a predictive model in machine learning algorithm predicts the yield of maize based on the input data. MSE is used to assess how accurately the model predicts the yield of maize based on the input variables in the context of maize yield prediction using ANN.

$$mse = \frac{1}{n}\sum_{i=1}^{n}((yi - ym))^2 - - - - - equ(3.4)$$

D. Root mean squared error

Root Mean Squared Error (RMSE) is used to assess how well a predictive model, predicts the yield of maize based on various inputs. The square root of the average of the squared differences between the predicted values and the actual values is what the RMSE[66]. RMSE is used to

assess how accurately the model predicts the yield of maize based on the input variables in the context of maize yield prediction using ANN. If the RMSE is lower, the model is more accurate at predicting maize yield; if it is greater, the model is less accurate.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \quad (yi - ym)^2}$$----------eq (3.5)

3.2.10. Implementation Tools

This study uses multiple implementation tools and packages to implement a proposed prediction, model. A Python-based programming language for putting each suggested fix into practice, from data preparation through model construction, as well as for reviewing implementations and potential classifier models. Because Python is the preferred programming language for researchers, developers, and data scientists who need to work with machine learning models, it is employed in this study. The implementation tools and Python packages used in this study are included in the table below, together with information on their versions and descriptions.

Table 3. 5: Description of the tools and python packages used during the implementation

No	Tool	Explanation
1	Python 3.9.0	A compile environment for python code easy to learn, powerful programming language to develop a machine learning application.
2	Subline text 3	Easy cross-platform code editor well-known for its speed, comfort of use. It's an incredible editor right out of the box, but the real power comes from the ability to enhance its functionality using Package Control and creating custom settings.
3	Microsoft Excel 2013	Used data preparation tasks in cleaning, filtering, sorting the collected data, and remove duplicated data Also, used to manage the explanation task.
4	Scikit-learn 0.24.2	A set of python modules for machine learning and data mining. This study uses it for feature extraction and training and testing model. The name of the package is called sklearn.
5	Pandas 1.2.4	High-performance, easy-to-use data structures, and data analysis tools. This study uses it for data reading, manipulation, writing and

		handling the data frame.
6	NumPy 1.20	Array processing for number, strings, and objects.
7	Matplotlib	It is a comprehensive plotting library in Python that provides a wide range of customizable plots and visualizations
8	Seaborn	Seaborn is a high-level data visualization library built on top of matplotlib and It provides a simplified interface to create attractive and informative statistical graphics.
9	TensorFlow	TensorFlow is an open-source library for numerical computation and machine learning
10	Keras	Keras is a high-level neural networks API that runs on top of TensorFlow.

Chapter Four

Result and Discussion

4.1. Introduction

This chapter focuses on the process of model implementation, evaluates the performance of the model, discuss about results of the model and the model discussion used to predict maize yield. Generally, the chapter discusses the dataset analysis, the model development, the model performance, the result analysis, and the model comparison. The results of the maize yield prediction using ensemble techniques by combining different machine learning algorithm like support vector machine, artificial neural network and decision tree would be demonstrated and discussed in this chapter. The objective of this chapter is to investigate and evaluate the manner in which these models performed in predicting maize yields using the data that was collected. The dataset that was used, how the models were created, how the findings were analyzed, and how the results from each model were compared are all covered in detail in the parts that follow.

4.2. Dataset Analysis

Preprocessing was performed on the dataset to standardize the data, get rid of outliers, and fill in any missing values. The dataset was then split into training and testing sets, with the training sets utilizing 80% of the data and the testing sets utilizing 20% of the data. The dataset analysis part of the paper should include a full explanation of the dataset that was used. This dataset contains historical data on maize yield together with other agricultural parameters like rainfall, temperature, and soil characteristics. It was collected from the Hadiya Zone of Ethiopia.

Various agronomic parameters and accompanying maize yield measurements make up the dataset used in this study. The dataset, which includes a wide variety of geographic regions, was gathered over numerous growing seasons. In order to verify that the dataset was of high quality and appropriate for training an effective maize yield prediction model, a thorough study of the dataset was done prior to model creation. In order to find patterns, distributions, and probable outliers in the dataset, the study involved data preprocessing, feature engineering, and exploratory data analysis.

4.2.1. Feature Selection

Table 4. 1: Correlation values of the parameter

Area/hectar	0.983915
Urea/kg	0.112434
Tmin	0.063428
Year	0.037806
Production	0.036727
Dap/kg	0.034160
Woreda	0.033241
hum	0.029887
RF	0.016912
Soil	0.016712
Tmax	0.013213
SS	0.001440

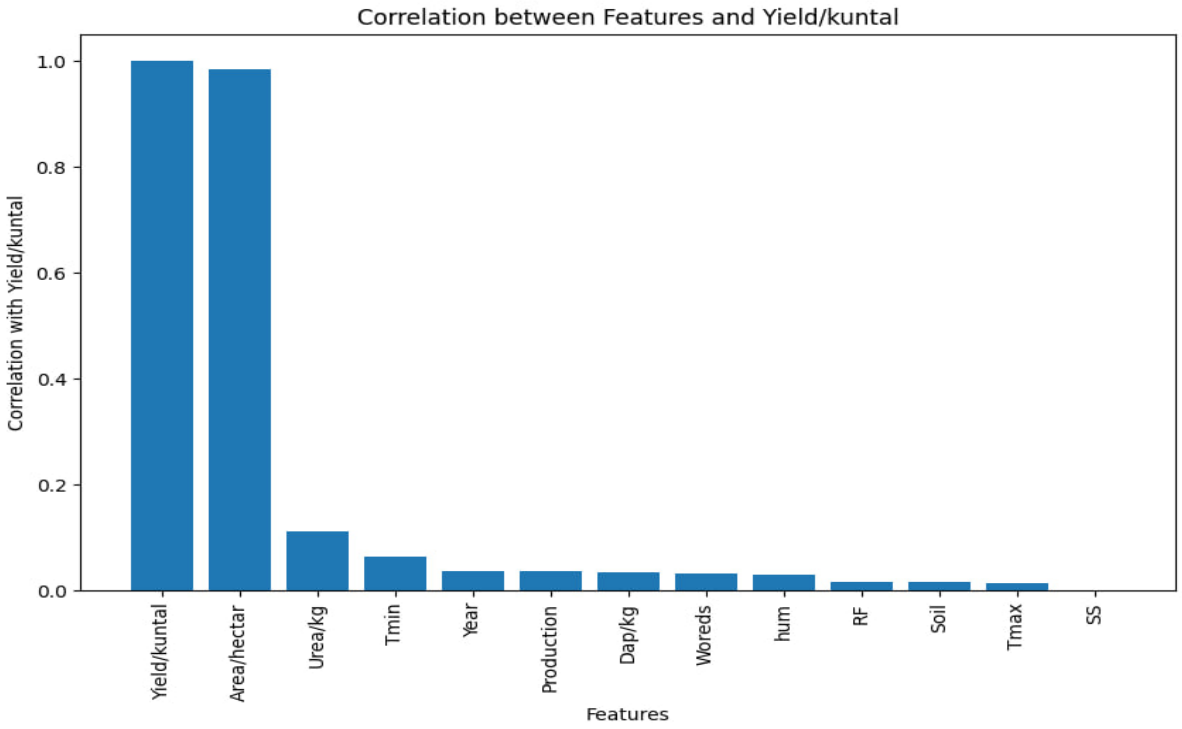

Figure 4. 1: The range of the correlation of the attribute respect to the yield of the crop

4.2.2. Exploratory data analysis

Exploratory data analysis (EDA), an important step in the machine learning process, is used to understand the properties of the dataset and gain insights into it before creating predictive

models. EDA helps with pattern identification, association discovery, outlier detection, and data quality assessment in the context of a dataset for maize yield prediction. We must use summary statistics (describe) to provide a broad overview of the central tendencies, dispersion, and distributions of the dataset. The quality of the data is assessed by using Isnull() function to identify missing values. Missing values need to be handled correctly in order for the following analysis to be valid and trustworthy. Data visualization, an essential part of EDA, is used to depict the distribution of the target variable, such as the yield of maize, and provides insight into its range. One illustration of how scatter plots can illustrate correlations between various variables. Examining these relationships is necessary to establish the relevance and potential predictive value of the input data. Correlation analysis, which is commonly displayed using a heatmap, can be used to comprehensively analyze the relationships between variables. The correlation heatmap visualizes the strength and direction of correlations, making it easier to choose characteristics and identify potential multicollinearity issues.

Table 4. 2: Exploratory data analysis for the dataset

	count	mean	std	min	25%
Year	4121.0	2012.428537	5.773374e+00	2003.000	2007.000000
Area/hectar	4121.0	13636.197835	7.933130e+04	0.050	49.000000
Production	4121.0	23.875295	1.139378e+01	2.000	20.000000
Urea/kg	4121.0	6.464689	2.714319e+02	0.050	0.334216
Dap/kg	4121.0	0.857383	4.437823e-01	0.125	0.821200
Soil	4121.0	1.000243	8.165461e-01	0.000	0.000000
Tmax	4121.0	19.811446	1.327237e+00	15.580	19.050000
Tmin	4121.0	7.120675	6.612810e+00	2.100	6.000000
RF	4121.0	124.822064	7.420124e+01	7.400	78.950000
hum	4121.0	76.974099	2.102785e+01	3.700	67.400000
SS	4121.0	5.857661	1.545812e+00	2.230	4.750000
Yield/kuntal	4121.0	320491.935600	1.881705e+06	0.650	1064.000000

	50%	75%	max
Year	2012.000000	2017.000000	2.022000e+03
Area/hectar	91.000000	916.000000	9.191919e+05
Production	24.000000	27.000000	2.870000e+02
Urea/kg	0.334216	0.334216	1.232300e+04

Dap/kg	0.821200	0.821200	3.575000e+00
Soil	1.000000	2.000000	2.000000e+00
Tmax	19.620000	20.450000	2.350000e+01
Tmin	7.120000	7.800000	8.200000e+01
RF	123.000000	168.060000	3.818000e+02
hum	71.700000	77.100000	1.540000e+02
SS	5.700000	7.000000	9.300000e+00
Yield/kuntal	2405.620000	25610.000000	2.663896e+07

From the above table, the statics are calculated for the column like year, area, production, fertilizer, soil, temperature, rainfall, sunshine and humidity, the count represent the number of non-null values in each columns, and the other statics represent the information about the distribution of the data.

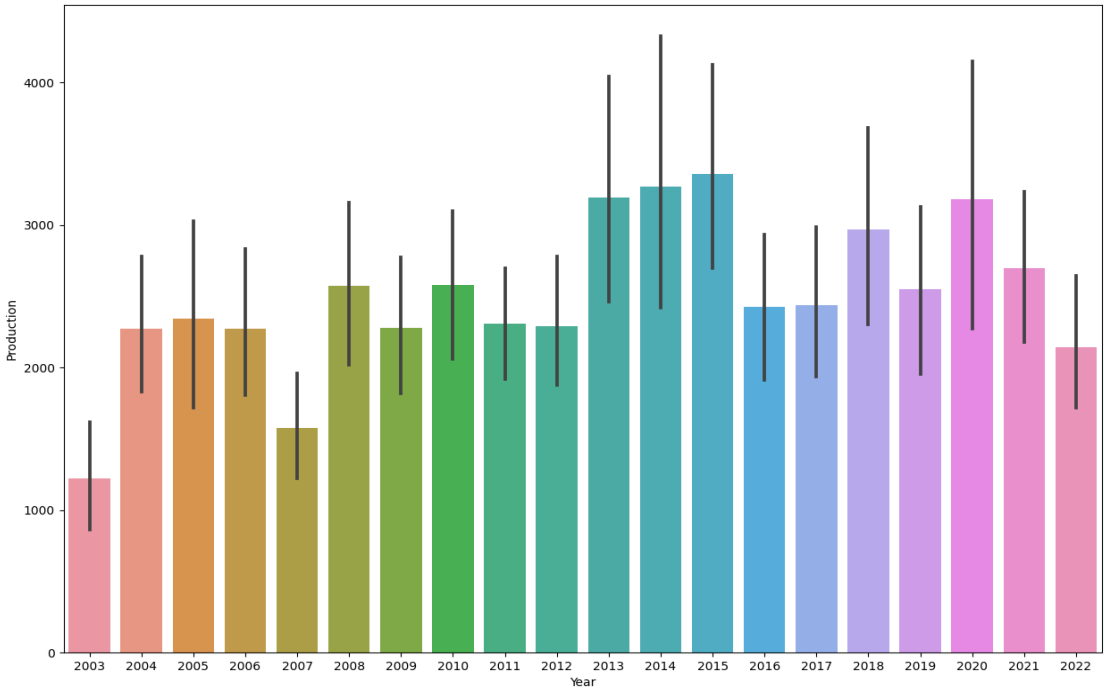

Figure 4. 2: The maize yield over the year

4.3. Model Development

With the use of the Python programming language and well-known learning frameworks like an artificial neural network (ANN), DVM and DT models was developed. To reduce the prediction error and improve the model's performance, the model was trained using a subset of the dataset and suitable training procedures such fully connected. For the development of model, we have used 825 testing datasets and 3296 training datasets for the prediction of maize yield.

4.3.1. Hyperparameters Selection

We have optimized the hyperparameters in order to get the best prediction performance. In this thesis, we have to use grid search tuning techniques for the optimization of parameters. For the development of the model that fits the data set, it needs to optimize the hyperparameters.

A. Artificial Neural Networks

In order to optimize the hyperparameters, the tuning process involves defining a param_grid dictionary that outlines the specific hyperparameters to be adjusted and the corresponding range of values to explore. These hyperparameters encompass various aspects such as the number of hidden layers, the number of neurons, the choice of activation function, the optimizer algorithm, the learning rate, the batch size, and the number of epochs. By specifying different values for each hyperparameter, a grid of potential combinations is generated.

In this setup, an Artificial Neural Network (ANN) is initialized with the Normal distribution for the kernel. The network's attributes, such as weights and learning rate, are adjusted using the Adam optimizer. The ANN consists of 16 neurons, which act as conduits for input information and generate output signals. The activation function used is Relu, which determines the level of neuron activity based on the independent features. The training process spans 100 epochs, representing the number of iterations over the training dataset, and utilizes a batch size of 10, indicating the number of training samples processed in a single iteration.

B. Support Vector Machine

In the case of the SVM model, the hyperparameters are C, gamma, and kernel. The parameter C controls the penalty for misclassifications on the training data and is set to 0.1. Gamma controls the shape of the decision boundary and is set to "scale," which means it is automatically

calculated based on the input data. The kernel is set to "linear," that determines the shape of the decision boundary and the mapping of the input data into a higher-dimensional feature space. The below graph show the hyperparameter tuning in support vector machine for each parameters of the model

```
SVM Best Parameters: {'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}
SVM RMSE: 0.035884136197916285
SVM R^2 Score: 0.719007212089611
Hyperparameter Selection Results:
```

	C	gamma	kernel	Mean Test Score (MSE)	R2 Score
0	0.1	scale	linear	0.001571	0.719007
1	0.1	scale	rbf	0.005304	0.719007
2	0.1	scale	poly	0.009261	0.719007
3	0.1	auto	linear	0.001571	0.719007
4	0.1	auto	rbf	0.005305	0.719007
5	0.1	auto	poly	0.009248	0.719007
6	1.0	scale	linear	0.001628	0.719007
7	1.0	scale	rbf	0.003422	0.719007
8	1.0	scale	poly	0.009261	0.719007
9	1.0	auto	linear	0.001628	0.719007
10	1.0	auto	rbf	0.003427	0.719007
11	1.0	auto	poly	0.009248	0.719007
12	10.0	scale	linear	0.004773	0.719007
13	10.0	scale	rbf	0.003422	0.719007
14	10.0	scale	poly	0.009261	0.719007
15	10.0	auto	linear	0.004773	0.719007
16	10.0	auto	rbf	0.003427	0.719007
17	10.0	auto	poly	0.009248	0.719007
18	100.0	scale	linear	0.303641	0.719007
19	100.0	scale	rbf	0.003422	0.719007
20	100.0	scale	poly	0.009261	0.719007
21	100.0	auto	linear	0.303641	0.719007
22	100.0	auto	rbf	0.003427	0.719007
23	100.0	auto	poly	0.009248	0.719007
24	200.0	scale	linear	1.212446	0.719007
25	200.0	scale	rbf	0.003422	0.719007
26	200.0	scale	poly	0.009261	0.719007
27	200.0	auto	linear	1.212446	0.719007
28	200.0	auto	rbf	0.003427	0.719007
29	200.0	auto	poly	0.009248	0.719007

Figure 4. 3: Hyperparameter selection for SVM

C. Decision Tree

For the DT model, the hyperparameters include max_depth, min_samples_split, and min_samples_leaf. Max_depth determines the maximum depth or levels of the decision tree and is set to 7. Min_samples_split sets the minimum number of samples required to split an internal node and is set to 5. Min_samples_leaf defines the minimum number of samples required to be in a leaf node and is set to 2.

```
DT performance: 0.9924811369182002
DT Best Parameters: {'max_depth': 7, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5}
Hyperparameter Selection Results:
    max_depth max_features  min_samples_leaf  min_samples_split  \
0           3         auto                 1                  2
1           3         auto                 1                  5
2           3         auto                 1                 10
3           3         auto                 2                  2
4           3         auto                 2                  5
..        ...          ...               ...                ...
76          7         log2                 2                  5
77          7         log2                 2                 10
78          7         log2                 4                  2
79          7         log2                 4                  5
80          7         log2                 4                 10

    Mean Test Score (MSE)  R2 Score
0                0.000102  0.997721
1                0.000106  0.997721
2                0.000112  0.997721
3                0.000106  0.997721
4                0.000106  0.997721
..                    ...       ...
76               0.001188  0.997721
77               0.003650  0.997721
78               0.002393  0.997721
79               0.002393  0.997721
80               0.002607  0.997721

[81 rows x 6 columns]
```

Figure 4. 4: Hyperparameter selection for DT

Table 4. 3: The Summary of Hyperparameters used for the model training

Model	Hyperparameters	Purpose	Values
ANN	Kernel initialize		Normal
	Optimizer	used to change the attributes of your neural network such as weights and learning rate	Adam
	Neuron	It is used for carrying input information and away outputs from the brain	16
	Activation function	How neuron is active based on the independent features	Relu
	Epochs	The number of iterations in the training set	100
	Bach size	Total number of training samples that are propagated in one iteration	10
SVM	C	controls the penalty for misclassifications on the training data	0.1
	Gamma	controls the shape of the decision boundary	Scale

	Kernel	Determines the shape of the decision boundary and the mapping of the input data into a higher-dimensional feature space.	Linear
DT	max_depth	determines the maximum depth or levels of the decision tree	7
	min_samples_split	sets the minimum number of samples required to split an internal node	5
	min_samples_leaf	defines the minimum number of samples required to be in a leaf node	2
	Criterion	specifies the measure used to evaluate the quality of a split	Entropy

In ANN, hyper parameters can be tuned by using the Keras regression function for grid search when developing the model. In ANN hyper parameter tuning, we have to consider the number of hidden layers in the model, the number of epochs, the number of batch sizes, the type of activation function, and the optimizer to use. In the ANN model, we optimize different parameters by using grid search hyper parameter tuning techniques. These parameters that we optimized are the number of neurons, epoch, batch size and activation function. In SVM, we have tuned C, gamma, and kernel, and in decision trees, we have used criteria, min_samples_leaf, max_depth, and min_samples_split hyper parameters.

The provided table outlines the hyperparameters and their corresponding values for three different models: artificial neural networks (ANN), support vector machines (SVM), and decision trees (DT). For the ANN model, the hyperparameters include kernel initialization, optimizer, and number of neurons, activation function, epochs, and batch size. The kernel is initialized using a normal distribution, the optimizer used is Adam, there are 16 neurons in the model, the activation function is set to Rectified linear unit (Relu), the training is performed for 100 epochs, and the batch size is set to 10.

These hyperparameters play a crucial role in model training and can significantly impact the model's performance and generalization capabilities. The specific values chosen for each

hyperparameter are based on considerations such as prior knowledge, experimentation, and optimization techniques to achieve the desired balance between under fitting and overfitting.

4.3.2. Experimental Results

A. Artificial neural network

The provided code encompasses several essential steps in training and evaluating a neural network model. Firstly, the model is compiled using the compile method, with the 'Adam' optimizer, 'MSE' loss function, and 'mean_absolute_error' metric. This sets the optimization algorithm, the type of loss to minimize, and the metric to monitor during training. Secondly, the model.summary() function generates a summary of the model's architecture, providing an overview of the number of parameters in each layer and the total number of trainable parameters. Following that, the model is trained using the fit method, where the training data (xtrain and ytrain) is supplied along with the batch size (10) and the number of epochs (100). The validation data (xtest, ytest) is employed to evaluate the model's performance on a separate dataset during the training process. Lastly, the training process yields a model_history object, which contains valuable information such as the loss and metric values recorded at each epoch. This information can be further analyzed and visualized to gain insights into the model's behavior and performance.

Table 4. 4: ANN Model development

Model: "sequential_2"

Layer (type)	Output Shape	Param #
dense_6 (Dense)	(None, 16)	208
dense_7 (Dense)	(None, 8)	136
dense_8 (Dense)	(None, 1)	9

Total params: 353 (1.38 KB)
Trainable params: 353 (1.38 KB)
Non-trainable params: 0 (0.00 Byte)

First the dataset was divided into training and testing sets. We then construct the model architecture. The Sequential class from Keras is used to define a sequential model. The add method is used to add layers to the model. Then, an output layer with a single unit and linear activation is implemented, followed by hidden layers with 16 neurons and Relu activation with single hidden layer. Compilation of the model, the compile method is used to create the model. Since binary classification jobs frequently employ it, the loss function is set to mean absolute error. Adam, a powerful optimization algorithm, is the optimizer's default setting. It is also possible to specify additional metrics, like the mean square error.

```
Epoch 1/100
103/103 [==============================] - 3s 10ms/step - loss: 0.0332 - mean_absolute_error: 0.0929 - val_loss: 4.7758e-05 - va
Epoch 2/100
103/103 [==============================] - 1s 8ms/step - loss: 3.5887e-04 - mean_absolute_error: 0.0041 - val_loss: 3.5386e-05 -
Epoch 3/100
103/103 [==============================] - 1s 5ms/step - loss: 3.4980e-04 - mean_absolute_error: 0.0031 - val_loss: 2.5120e-05 -
Epoch 4/100
103/103 [==============================] - 1s 6ms/step - loss: 3.4433e-04 - mean_absolute_error: 0.0022 - val_loss: 2.1769e-05 -
Epoch 5/100
103/103 [==============================] - 1s 8ms/step - loss: 3.4338e-04 - mean_absolute_error: 0.0018 - val_loss: 2.1182e-05 -
Epoch 6/100
103/103 [==============================] - 1s 6ms/step - loss: 3.4322e-04 - mean_absolute_error: 0.0017 - val_loss: 2.1022e-05 -
Epoch 7/100
103/103 [==============================] - 1s 5ms/step - loss: 3.4316e-04 - mean_absolute_error: 0.0016 - val_loss: 2.0952e-05 -
Epoch 8/100
103/103 [==============================] - 0s 5ms/step - loss: 3.4314e-04 - mean_absolute_error: 0.0016 - val_loss: 2.0912e-05 -
Epoch 9/100
103/103 [==============================] - 1s 7ms/step - loss: 3.4311e-04 - mean_absolute_error: 0.0016 - val_loss: 2.0893e-05 -
Epoch 10/100
103/103 [==============================] - 1s 5ms/step - loss: 3.4309e-04 - mean_absolute_error: 0.0016 - val_loss: 2.0875e-05 -
Epoch 11/100
103/103 [==============================] - 0s 5ms/step - loss: 3.4306e-04 - mean_absolute_error: 0.0016 - val_loss: 2.0864e-05 -
Epoch 12/100
103/103 [==============================] - 1s 6ms/step - loss: 3.4303e-04 - mean_absolute_error: 0.0016 - val_loss: 2.0841e-05 -
Epoch 13/100
103/103 [==============================] - 1s 5ms/step - loss: 3.4298e-04 - mean_absolute_error: 0.0016 - val_loss: 2.0837e-05 -
```

Figure 4. 5: The number of Iteration in ANN for Maize yield prediction

The graph used to represent the number of iterations per epoch, which pass forward and backward during the model training, If the number of epochs increases, the prediction error decreases when we train the ANN model for the prediction of maize yield. Then it is necessary to train the model. Using the training set of data, the model is trained using the fit approach. The epoch's option controls how many times the model runs across the entire training dataset. The number of samples that must be processed before the internal model parameters can be altered is specified by the batch_size option. Finally, evaluate the model. Utilizing the evaluate method on the test data (X_test and Y_test) after training, the model may be evaluated. The method returns

the loss value and any specified metrics, and then prints the result. The result is described in the below section.

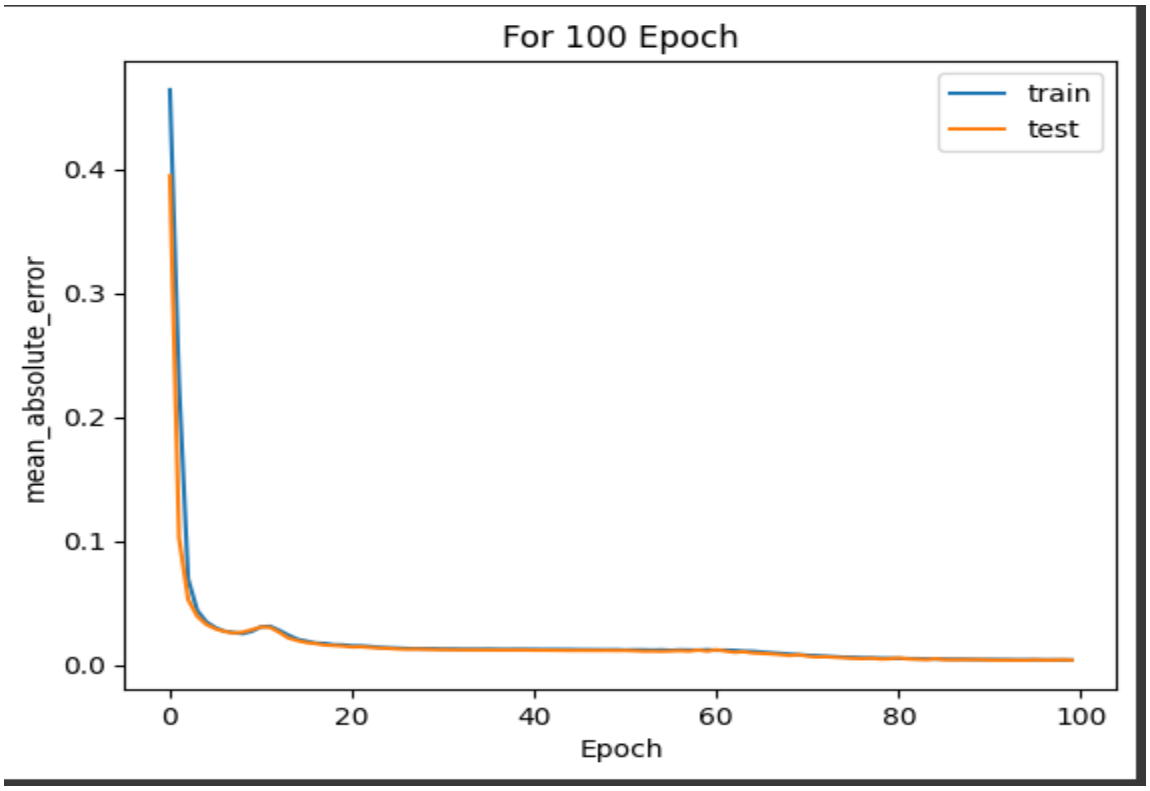

Figure 4. 6: Train test error over the number of iterations

We create a Sequential model and add layers using the dense class from Keras. We specify the size of the hidden layer using the hidden_layer_size variable, the number of neurons in the input layer using the input_layer_neurons variable, and the number of neurons in the output layer using the output_layer_neurons variable. The output layer does not have an explicit activation function specified, resulting in a linear activation by default.

B. Support vector machine

An SVM model is initialized with the SVR class, specifying the linear kernel, a regularization parameter C of 0.1, and the 'scale' option for gamma. The Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared score, and Root Mean Squared Error (RMSE) are calculated using appropriate functions from sklearn and math libraries. Overall, the SVM model demonstrates the training, prediction, and evaluation process for an SVM regression model using

the linear kernel. The evaluation metrics provide insights into the performance of the SVM model in predicting the target variable.

C. Decision Tree

The decision tree model is configured with specific hyperparameters such as max_depth=7 to limit the tree's depth, min_samples_split=5 to set the minimum number of samples required to split a node, and min_samples_leaf=2 to determine the minimum number of samples at a leaf node. The model is then trained the training datasets. Subsequently, the trained decision tree model is used to make predictions on the test set dataset. The specified hyperparameters control the tree's complexity and behavior, while the predicted values can be further evaluated and compared against the actual values to assess the model's performance.

4.4. Model Performance

The accuracy, dependability, and generalization capacities of the developed Ensemble model were assessed using a variety of measures. The metrics employed could be coefficient of determination (R-squared), mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), and mean squared error (MSE). To reduce overfitting and offer a reliable estimate of the model's performance, hyperparameter tuning were used. Additionally, to examine the correlation between projected and actual maize yield figures, visualizations like scatter plot were used.

An artificial neural network is built for the prediction of maize yield and trains the dataset. The training of the ANN can be performed using a batch size of 10 and epochs of 100. The result can suggest the Adam optimization algorithm with uniform weight initialization and the sigmoid activation function. And the best performance result can be achieved with a network number of 16 in the hidden layer and the following performance metrics: The performance of the prediction of yield by using ANN is 0.0001 MSE, 0.0061 MAE, 0.0122 RMSE, and 0.9702 % of the coefficient of determination. For the development of the SVM mode, the researcher uses C of 0.1, a kernel of linearity, and a scale of gamma. The research uses the SVM model for the prediction of the maize yield and has a performance of 0.0012 MSE, 0.0281 MAE, 0.0357 RMSE, and 0.8020% of the coefficient of determination.

And also, the research used the DT model for the prediction of the maize yield. For the development of the DT mode, the researcher used criteria of entropy: Maximum Tree Depth of 7, Minimum Number of Samples split of 5 and Minimum Number of Samples in a Leaf Node of 2, with a performance of 0.0034 MSE, 0. 0061 MAE, 0.0058 RMSE, and 0. 9924% of the coefficient of determination. In this thesis, the artificial neural network (ANN), decision tree (DT), and SVM model are combined using the stacking ensemble technique. The stacking ensemble model takes advantage of multiple base models to improve performance predictions.

This research uses the stacking ensemble technique for the prediction of the maize yield. After the development of the stacking ensemble technique mode got the performance of 0. 0032 MSE, 0. 0025 MAE, 0. 0057 RMSE, and 0. 9928 of the coefficient of determination.

Table 4. 5: Performance comparison among algorithms

Model	MAE	MSE	RMSE	R2_score
ANN	0. 0061	0. 0001	0.0122	0. 9702
SVM	0. 0281	0. 0012	0. 0357	0. 8020
DT	0. 0061	0. 0034	0. 0058	0. 9924
Ensemble	0. 0025	0.0032	0.0057	0.9928

The most crucial data processing techniques for enhancing maize yield prediction performance are data normalization, feature selection, and handling missing data. In this study, we employed the min-max normalization method to ensure that all input variables were on a similar scale, preventing any individual feature from dominating the prediction process. For feature selection, we utilized correlation analysis to identify and remove irrelevant columns, such as the crop column, thereby improving the generalization of our model. Additionally, Principal Component Analysis (PCA) was employed for feature extraction.

Comparing the performance of the model with and without data processing techniques, it was evident that utilizing preprocessed data yielded significantly better results. Without data processing, the performance metrics included a mean squared error (MSE) value of 0.2445, a mean absolute error (MAE) value of 0.42787, an R2 score of 0.928, and a root mean squared error (RMSE) value of 0.5636. In contrast, the normalized (scaled) data exhibited improved performance, with an MSE value of 0.0032, an MAE value of 0.0025, an RMSE value of 0.0057,

and a coefficient of determination of 0.9928. Similarly, when feature selection was applied, the performance metrics remained consistent with the normalized data. Interestingly, the ensemble model performed better when using the normalized dataset. Therefore, it is advisable to assess the accuracy parameter for both raw and normalized data to ensure a comprehensive analysis.

4.5. Comparison with Existing Models

The resulting ensemble model and other models or traditional methods for forecasting maize production were compared. These could consist of machine learning methods, statistical regression models, or domain-specific models. Prediction accuracy, computational efficiency, and ease of implementation were taken into consideration while comparing the performance metrics of the ensemble model to those of other models including SVM, ANN, and DT. This stage sought to emphasize the benefits and enhancements provided by the ensemble-based method for predicting maize prediction is 0.0032 MSE, 0. 0025 MAE, 0.0057 RMSE, and 0.9928 % of the coefficient of determination.

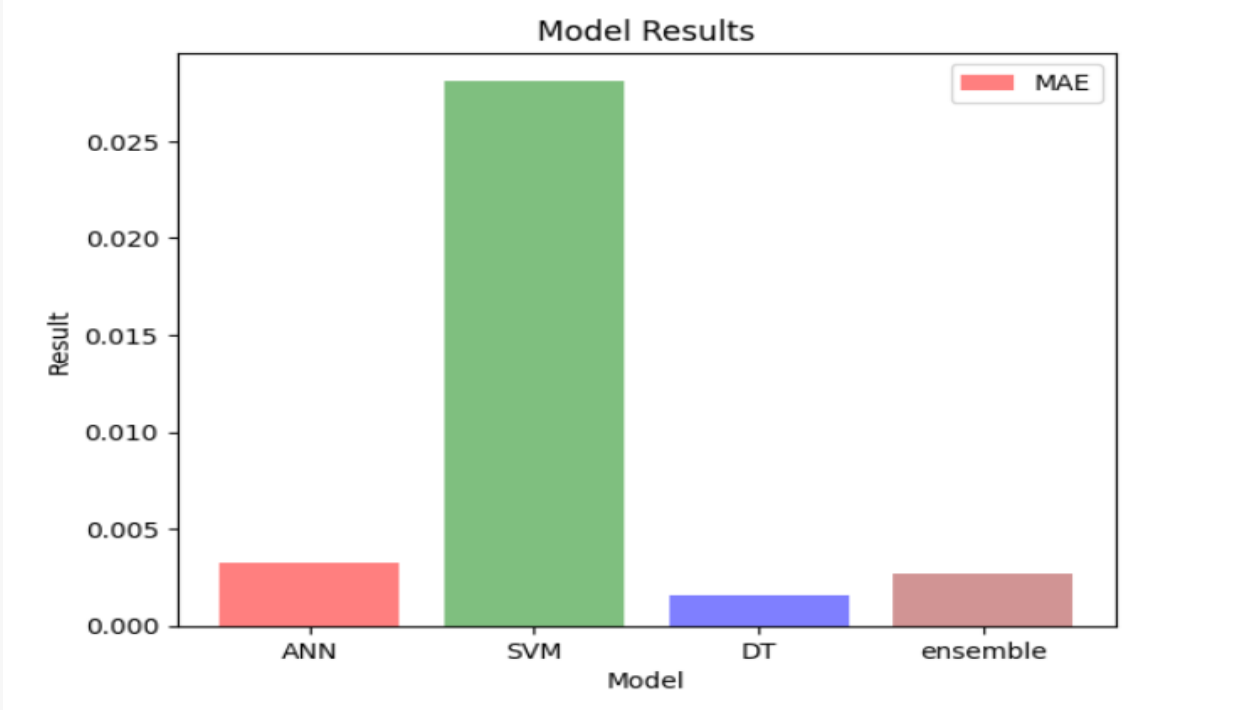

Figure 4. 7: The Comparison of model performance of MAE

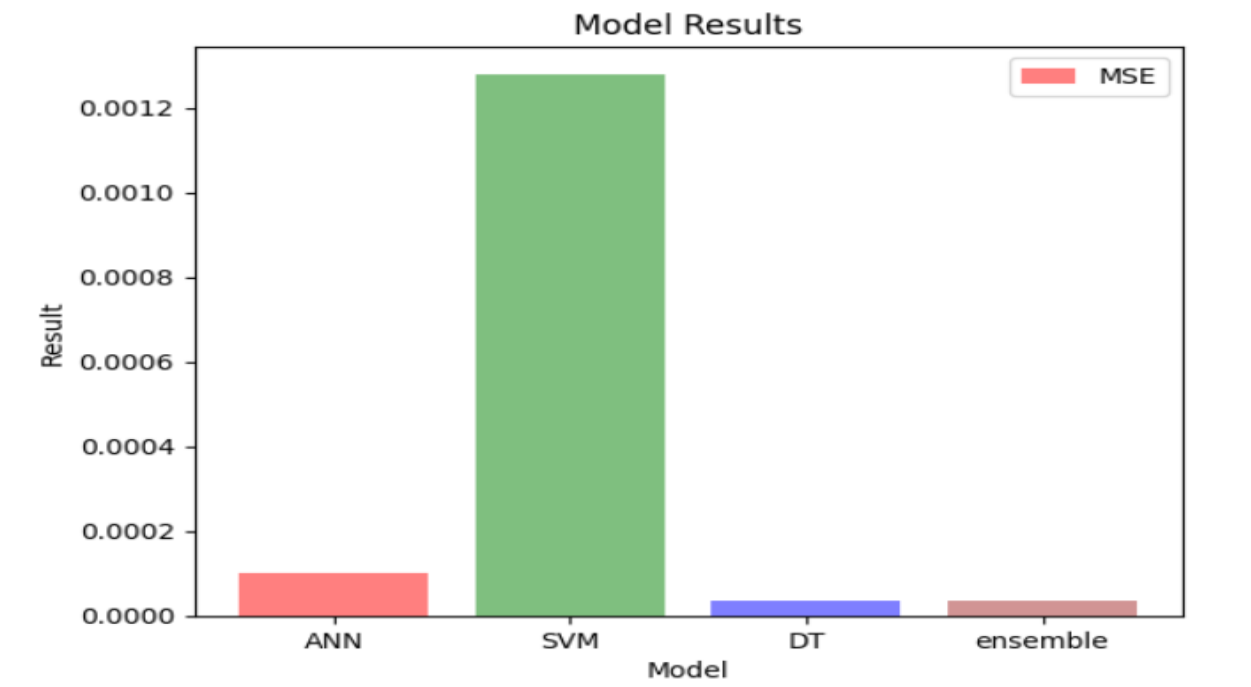

Figure 4. 8: The Comparison of model performance of MSE

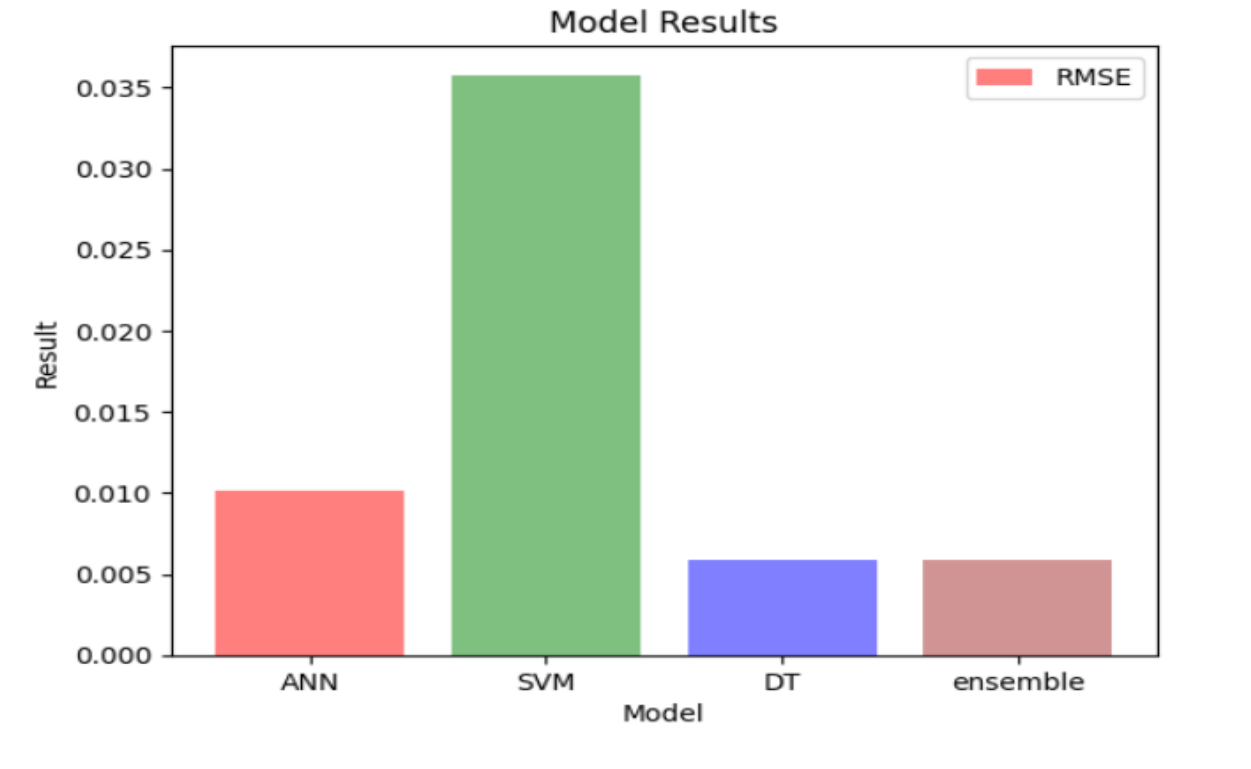

Figure 4. 9: The Comparison of model performance of RMSE

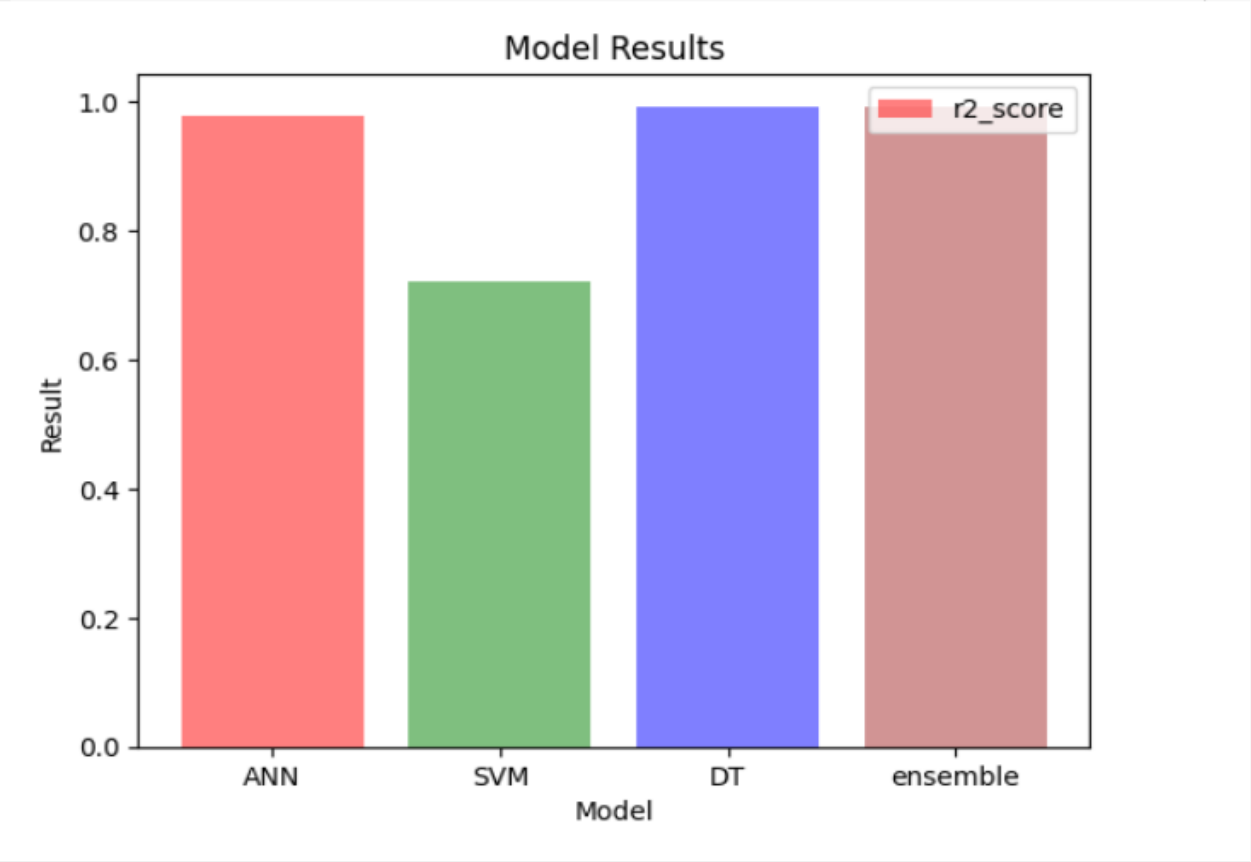

Figure 4. 10: The Comparison of model performance of R2_score

Table 4. 6: Model comparison with the existing model

Paper	Ref	Crop Type	Parameters	Model	Performance			
					MSE	RMSE	MAE	R^2
Existing Paper	[30]	Maize	gridded biophysical and socio-economic variables	RF	-	-	-	0.63
	[31]	Maize	satellite imaging, meteorological data, and soil data	SVMl, and NNET		0.13		

	[45]	Soybean	Area , previous yield history	ANN	3.915	-	1.813	0.94
	[36]	six crops	meteorological and soil data	DT, KNN, and LR	-	-	-	0.95
					MSE	RMSE	MAE	R^2
Proposed Work	maize yield prediction		Year, Kebele, Area, Urea, Dap, Class, Tmax, Tmin, RF, hum, and SS, yield	ensemble	0. 0032	0. 0057	0. 0025	0. 9928

4.6. Result Analysis

We carefully looked over and evaluated the outcomes of the ensemble model's forecast of maize yield. In the analysis, the most significant agronomic factors that influence forecasts of maize production were identified, and the effects of each factor on prediction accuracy were assessed. Any limitations or assumptions made throughout the model construction process were also examined. The significance of the findings for agricultural practices and their potential application in crop management and decision-making were taken into account.

Finally, recommendations for more research or model enhancements were offered to advance upcoming breakthroughs in maize yield prediction. We assume that the test error can be retrieved using the data set aside for testing, while the training error can be obtained using the same data that the model was trained with. The test error demonstrates how effectively the network generalizes to new data samples whereas the training error reflects how well the model fits the data. The mean absolute error, mean square error, root mean square error, coefficient of determination, and accuracy have all been used to gauge the model's performance. Therefore, the performance evaluation of the ensemble techniques for predicting maize yield is 0.0032 MSE, 0. 0025 MAE, 0. 0057 RMSE, and 0. 9928% of the coefficient of determination. The objective of this analysis is to predict the yield of crops based on the crops cultivated in each area with

different parameters. By predicting the potential yield for maize in each area, we aim to provide valuable insights to farmers with the highest yield for future cultivation. This analysis aims to optimize farmers' benefits by maximizing their agricultural output.

Chapter Five

Conclusions and Recommendation

5.1. Conclusions

The machine learning system is driven to make the best predictions possible and has the best option for forecasting maize output. In order to forecast maize production, many machine learning techniques were discussed in the research. The document is used to address challenges that Ethiopian farmers, particularly in the Hadiya Zone, have while analyzing and forecasting the maize production data to farm their properties. Machine learning approaches, including artificial neural networks (ANNs), support vector machines (SVM), decision trees (DT), and ensemble techniques, were used in this study to forecast maize yield. The goal of the study was to gather data that would aid in the planning, management, and resource allocation decisions that farmers and policymakers would need to make. The research's findings show that the ensemble approach model is more accurate in its capacity to forecast maize production. The model's performance was assessed using a number of performance indicators, including mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R2). Due to the low MAE and RMSE values and strong R2 value of the ensemble approach model, maize yield could be forecasted. This study illustrated how ensemble techniques in machine learning, in particular, can be used to forecast maize yield. The study seeks to increase local maize production while also adding to the body of knowledge on the application of machine learning in agriculture. The collection includes 4121 records with various parameters relating to weather, soil, and maize yield information. Various methods were utilized to create and evaluate the models based on the variables affecting the maize yield. Because they quantify the best performance, ensemble approaches are the best model for predicting maize yield.

5.2. Future Works

The research's findings show that the ensemble approach model is more accurate in its capacity to forecast maize production. The model's performance was assessed using a number of performance indicators, including mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R2). Due to the low MAE and RMSE values and strong R2 value of the ensemble approach model, maize yield could be forecasted. This study illustrated how ensemble techniques in machine learning, in particular, can be used to forecast maize yield. The study seeks to increase local maize production while also adding to the body of knowledge on the application of machine learning in agriculture. We may analyze the elements that are utilized to influence maize production that bring a greater yield and a higher profit within the same cost by using the dataset of this study and some other additional datasets. According to the cost of the crop, the report advised the researcher to choose a crop and fertilizer. The researcher suggested that it could be preferable to expand the study by including further criteria like the price and the contribution from exports.

References

[1] G. Leta, G. Kelboro, T. Stellmacher, and A.-K. Hornidge, “The agricultural extension system in Ethiopia: operational setup, challenges and opportunities,” 2017.

[2] D. Alemu and A. Assaye, “The Political Economy of the Rice Value Chain in Ethiopia: Actors, Performance, and Discourses,” 2021.

[3] D. Alemu and K. Berhanu, “The Political Economy of Agricultural Commercialisation in Ethiopia: Discourses, Actors and Structural Impediments,” 2018.

[4] M. Khan, B. Jan, H. Farman, J. Ahmad, H. Farman, and Z. Jan, “Deep learning methods and applications,” *Deep Learn. Converg. Big Data Anal.*, pp. 31–42, 2019.

[5] L. Deng and D. Yu, “Deep learning: methods and applications,” *Found. Trends® Signal Process.*, vol. 7, no. 3–4, pp. 197–387, 2014.

[6] M. Awad and R. Khanna, “Support vector regression,” in *Efficient learning machines*, Springer, 2015, pp. 67–80.

[7] A. Kumar, N. Kumar, and V. Vats, “Efficient crop yield prediction using machine learning algorithms,” *Int. Res. J. Eng. Technol.*, vol. 5, no. 06, pp. 3151–3159, 2018.

[8] M. Shahhosseini, G. Hu, S. Khaki, and S. V. Archontoulis, “Corn yield prediction with ensemble CNN-DNN,” *Front. Plant Sci.*, vol. 12, p. 709008, 2021.

[9] S. B. Awulachew, “Irrigation potential in Ethiopia: Constraints and opportunities for enhancing the system,” *Gates Open Res*, vol. 3, no. 22, p. 22, 2019.

[10] D. Lee *et al.*, “Maize yield forecasts for Sub-Saharan Africa using Earth Observation data and machine learning,” *Glob. Food Secur.*, vol. 33, p. 100643, 2022.

[11] M. Shahhosseini, R. A. Martinez-Feria, G. Hu, and S. V. Archontoulis, “Maize yield and nitrate loss prediction with machine learning algorithms,” *Environ. Res. Lett.*, vol. 14, no. 12, p. 124026, 2019.

[12] Seleshi Bekele Awulachew, “Irrigation potential in Ethiopia: constraints and opportunities for enhancing the system,” 2019, doi: 10.21955/GATESOPENRES.1114943.1.

[13] D. Biru and J. Tefera, “Maize Yield Forecast Using GIS and Remote Sensing The Case of Kaffa Zone, South Western Ethiopia,” 2021.

[14] A. Dawit and B. Kassahun, “The political economy of agricultural commercialisation in Ethiopia: discourses, actors and structural impediments.,” *Work. Pap.-Agric. Policy Res. Afr. APRA*, no. 14, 2018.

[15] B. Merga and J. Haji, “Factors impeding effective crop production in Ethiopia,” *J. Agric. Sci.*, vol. 11, no. 10, pp. 1–14, 2019.

[16] D. A. Bondre and Mahagaonkar and S. Mahagaonkar, “Prediction of crop yield and fertilizer recommendation using machine learning algorithms,” *Int. J. Eng. Appl. Sci. Technol.*, vol. 4, no. 5, pp. 371–376, 2019.

[17] K. Chaudhary and F. Kausar, “PREDICTION OF CROP YIELD USING MACHINE LEARNING”.

[18] K. Mulatu, G. Bogale, B. Tolesa, M. Worku, Y. Desalegne, and A. Afeta, “Maize production trends and research in Ethiopia,” in *1. National Maize Workshop of Ethiopia, Addis Abeba (Ethiopia), 5-7 May 1992*, IAR, 1993.

[19] M. Worku, H. Tuna, M. Nigussie, A. Deressa, D. Tanner, and S. Twumasi-Afriyie, “Maize production trends and research in Ethiopia,” *Mand. N Tann. DG Twumasi-Afriyie Seds*, pp. 10–14, 2002.

[20] T. Banavlikar, A. Mahir, M. Budukh, and S. Dhodapkar, “Crop recommendation system using Neural Networks,” *Int. Res. J. Eng. Technol. IRJET*, vol. 5, no. 5, pp. 1475–1480, 2018.

[21] T. O. Ayodele, “Types of machine learning algorithms,” *New Adv. Mach. Learn.*, vol. 3, pp. 19–48, 2010.

[22] D. A. Bondre and S. Mahagaonkar, “Prediction of crop yield and fertilizer recommendation using machine learning algorithms,” *Int. J. Eng. Appl. Sci. Technol.*, vol. 4, no. 5, pp. 371–376, 2019.

[23] H. E. Siddharth, M. Bharathi, G. Navya, P. Kumar, and H. Doddamani, “Machine learning techniques for selecting the crop to increase the yield,” 2019.

[24] M. Mathapathi, K. K. Patil, G. Manjushree, and K. N. NM, “Predicting Yield of Crop and Detecting Fertilizer Efficiency”.

[25] B. Charbuty and A. Abdulazeez, “Classification based on decision tree algorithm for machine learning,” *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[26] P. Cunningham, “Ensemble techniques,” *Techn Own Port Scan Dataset Shown Fig.*, vol. 6, 2007.

[27] M. Keerthana, K. J. M. Meghana, S. Pravallika, and M. Kavitha, “An ensemble algorithm for crop yield prediction,” in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, IEEE, 2021, pp. 963–970.

[28] K. Archana and K. G. Saranya, “Crop yield prediction, forecasting and fertilizer recommendation using voting based ensemble classifier,” *SSRG Int J Comput Sci Eng*, vol. 7, pp. 1–4, 2020.

[29] A. N. Mulatu and E. Tamir, “Prediction of Teff Yield Using a Machine Learning Approach,” in *Artificial Intelligence and Digitalization for Sustainable Development: 10th EAI International Conference, ICAST 2022, Bahir Dar, Ethiopia, November 4-6, 2022, Proceedings*, Springer, 2023, pp. 159–176.

[30] F. Muthoni, C. Thierfelder, B. Mudereri, J. Manda, M. Bekunda, and I. Hoeschle-Zeledon, “Machine learning model accurately predict maize grain yields in conservation agriculture systems in Southern Africa,” in *2021 9th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, IEEE, 2021, pp. 1–5.

[31] M. Croci, G. Impollonia, M. Meroni, and S. Amaducci, “Dynamic Maize Yield Predictions Using Machine Learning on Multi-Source Data,” *Remote Sens.*, vol. 15, no. 1, p. 100, 2022.

[32] M. Shahhosseini, R. A. Martinez-Feria, G. Hu, and S. V. Archontoulis, “Maize yield and nitrate loss prediction with machine learning algorithms,” *Environ. Res. Lett.*, vol. 14, no. 12, p. 124026, 2019.

[33] F. H. R. Baio *et al.*, “Maize Yield Prediction with Machine Learning, Spectral Variables and Irrigation Management,” *Remote Sens.*, vol. 15, no. 1, p. 79, 2022.

[34] D. Biru and J. Tefera, “Maize Yield Forecast Using GIS and Remote Sensing The Case of Kaffa Zone, South Western Ethiopia,” 2021.

[35] N. Gandhi, O. Petkar, and L. J. Armstrong, “Rice crop yield prediction using artificial neural networks,” in *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, IEEE, 2016, pp. 105–110.

[36] D. J. Reddy and M. R. Kumar, “Crop yield prediction using machine learning algorithm,” in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2021, pp. 1466–1470.

[37] W. Mupangwa, L. Chipindu, I. Nyagumbo, S. Mkuhlani, and G. Sisito, “Evaluating machine learning algorithms for predicting maize yield under conservation agriculture in Eastern and Southern Africa,” *SN Appl. Sci.*, vol. 2, pp. 1–14, 2020.

[38] C. Dang, Y. Liu, H. Yue, J. Qian, and R. Zhu, “Autumn crop yield prediction using data-driven approaches:-support vector machines, random forest, and deep neural network methods,” *Can. J. Remote Sens.*, vol. 47, no. 2, pp. 162–181, 2021.

[39] A. A. Tesfaye, B. G. Awoke, T. S. Sida, and D. E. Osgood, “Enhancing smallholder wheat yield prediction through sensor fusion and phenology with machine learning and deep learning methods,” *Agriculture*, vol. 12, no. 9, p. 1352, 2022.

[40] L. S. Cedric *et al.*, “Crops yield prediction based on machine learning models: Case of West African countries,” *Smart Agric. Technol.*, vol. 2, p. 100049, 2022.

[41] A. Nigam, S. Garg, A. Agrawal, and P. Agrawal, “Crop yield prediction using machine learning algorithms,” in *2019 Fifth International Conference on Image Information Processing (ICIIP)*, IEEE, 2019, pp. 125–130.

[42] D. A. Bondre and S. Mahagaonkar, “Prediction of crop yield and fertilizer recommendation using machine learning algorithms,” *Int. J. Eng. Appl. Sci. Technol.*, vol. 4, no. 5, pp. 371–376, 2019.

[43] F. Siva, “Smart fertilizer recommendation through NPK analysis using Artificial Neural Networks,” Strathmore University, 2019.

[44] W. W. Guo and H. Xue, “Crop yield forecasting using artificial neural networks: A comparison between spatial and temporal models,” *Math. Probl. Eng.*, vol. 2014, 2014.

[45] E. R. Abraham *et al.*, “Time series prediction with artificial neural networks: An analysis using Brazilian soybean production,” *Agriculture*, vol. 10, no. 10, p. 475, 2020.

[46] A. Yadav, C. K. Jha, and A. Sharan, “Optimizing LSTM for time series prediction in Indian stock market,” *Procedia Comput. Sci.*, vol. 167, pp. 2091–2100, 2020.

[47] R. Medar, V. S. Rajpurohit, and S. Shweta, “Crop yield prediction using machine learning techniques,” in *2019 IEEE 5th international conference for convergence in technology (I2CT)*, IEEE, 2019, pp. 1–5.

[48] M. B. Santosh Kumar and K. Balakrishnan, “Development of a model recommender system for agriculture using apriori algorithm,” in *Cognitive Informatics and Soft Computing: Proceeding of CISC 2017*, Springer, 2019, pp. 153–163.

[49] A. Nigam, S. Garg, A. Agrawal, and P. Agrawal, “Crop yield prediction using machine learning algorithms,” in *2019 Fifth International Conference on Image Information Processing (ICIIP)*, IEEE, 2019, pp. 125–130.

[50] R. Medar, V. S. Rajpurohit, and S. Shweta, “Crop yield prediction using machine learning techniques,” in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, IEEE, 2019, pp. 1–5.

[51] Y. J. N. Kumar, V. Spandana, V. S. Vaishnavi, K. Neha, and V. Devi, “Supervised machine learning approach for crop yield prediction in agriculture sector,” in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2020, pp. 736–741.

[52] N. Gandhi, O. Petkar, and L. J. Armstrong, “Rice crop yield prediction using artificial neural networks,” in *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, IEEE, 2016, pp. 105–110.

[53] S. Ujjainia, P. Gautam, and S. Veenadhari, “A Crop Recommendation System to Improve Crop Productivity using Ensemble Technique,” *Int. J. Innov. Technol. Explor. Eng. IJITEE*, vol. 10, no. 4, 2021.

[54] M. Shahhosseini, G. Hu, I. Huber, and S. V. Archontoulis, “Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt,” *Sci. Rep.*, vol. 11, no. 1, p. 1606, 2021.

[55] S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika, and J. Nisha, “Crop recommendation system for precision agriculture,” in *2016 Eighth International Conference on Advanced Computing (ICoAC)*, IEEE, 2017, pp. 32–36.

[56] N. H. Kulkarni, G. N. Srinivasan, B. M. Sagar, and N. K. Cauvery, “Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique,” in *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, IEEE, 2018, pp. 114–119.

[57] Y. J. N. Kumar, V. Spandana, V. S. Vaishnavi, K. Neha, and V. Devi, “Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector,” in *2020 5th*

International Conference on Communication and Electronics Systems (ICCES), IEEE, 2020, pp. 736–741.

[58] R. K. Rajak, A. Pawar, M. Pendke, P. Shinde, S. Rathod, and A. Devare, "Crop recommendation system to maximize crop yield using machine learning technique," *Int. Res. J. Eng. Technol.*, vol. 4, no. 12, pp. 950–953, 2017.

[59] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 111–117, 2006.

[60] T. Ahmad and M. N. Aziz, "Data preprocessing and feature selection for machine learning intrusion detection systems," *ICIC Express Lett*, vol. 13, no. 2, pp. 93–101, 2019.

[61] C. V. G. Zelaya, "Towards explaining the effects of data preprocessing on machine learning," in *2019 IEEE 35th international conference on data engineering (ICDE)*, IEEE, 2019, pp. 2086–2090.

[62] C. Liu, *Data transformation: Standardization vs normalization*. KDnuggets, 2022.

[63] S. Patro and K. K. Sahu, "Normalization: A preprocessing stage," *ArXiv Prepr. ArXiv150306462*, 2015.

[64] D. J. Ozer, "Correlation and the coefficient of determination.," *Psychol. Bull.*, vol. 97, no. 2, p. 307, 1985.

[65] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)," *Geosci. Model Dev. Discuss.*, vol. 7, no. 1, pp. 1525–1534, 2014.

[66] M. Abedinpour, A. Sarangi, T. B. S. Rajput, M. Singh, H. Pathak, and T. Ahmad, "Performance evaluation of AquaCrop model for maize crop in a semi-arid environment," *Agric. Water Manag.*, vol. 110, pp. 55–66, 2012.