HOW MACHINE LEARNING CAN EVALUATE THE INFLUENCE OF SOCIOECONOMIC AND CLIMATIC FACTORS IN AGRICULTURAL YIELD: A CASE OF NIGERIA

A dissertation presented to the department of Computer Science,

African University of Science and Technology, Abuja-Nigeria

In partial fulfilment of the requirements for a Masters degree in Management of Information Technology

By

Dappa Tamuno-Opubo Godwin (41051)

Abuja, Nigeria

MAY, 2023

CERTIFICATION

This is to certify that the thesis titled HOW MACHINE LEARNING CAN EVALUATE THE INFLUENCE OF SOCIOECONOMIC AND CLIMATIC FACTORS IN AGRICULTURAL YIELD: A CASE OF NIGERIA submitted to the school of postgraduate studies, African University of Science and Technology (AUST), Abuja, Nigeria for the award of the Master's degree is a record of original research carried out by Dappa Tamuno-Opubo Godwin in the Department of Computer Science.

07.05.2023

SIGNATURE PAGE

HOW MACHINE LEARNING CAN EVALUATE THE INFLUENCE OF SOCIOECONOMIC AND CLIMATIC FACTORS IN AGRICULTURAL YIELD: A CASE OF NIGERIA

By

Dappa Tamuno – Opubo Godwin

A THESIS APPROVED BY THE COMPUTER SCIENCE DEPARTMENT

RECOMMENDED: _____

Supervisor: Dr. Rajesh Prasad

_____

Head of Department: Dr. Rajesh Prasad

APPROVED: _____

Chief Academic Officer

_____

Date

ABSTRACT

The major international agencies in charge of nutrition are becoming increasingly concerned about global agricultural production in particular. Food insecurity has emerged in some populated areas, including Africa, as a result of the increased worldwide need for food as a result of record population growth. Climate change and its variability are two additional factors that contribute to world food insecurity. Furthermore, agricultural policy officials, farmers, and decision-makers require advanced technologies in order to make timely strategies or policies that will have an effect on the quality of crop harvests. Machine learning and other new, powerful analytical techniques made possible by big data technologies have already proven useful in a number of industries, including biology, finance, and medicine. The yield of three major crops, including cocoa, sesame, and cashew, at the national level in Nigeria during the course of the years spanning 1990 to 2020 is forecasted in this study using a machine learning-based prediction method. We used climatic, agricultural yield, and socioeconomic data to help policymakers and farmers anticipate the yearly agricultural output in Nigeria. We employed k-nearest neighbors, a decision tree, and random forest. We also employed a hyper-parameter tweaking technique through cross-validation to enhance the model and avoid overfitting. For sesame, the accuracy of the Decision Tree model was the highest, having a test accuracy of 97.92% for socioeconomic and climatic factors combined, while the KNN model did the best with a test accuracy of 99.71% for climatic components separately. The accuracy of the Random Forest model was 87.54% for climatic elements alone and 87.64% for socioeconomic and economic factors together. For cocoa, the Decision Tree model had an accuracy of 89.49% for socioeconomic and climatic factors combined and 89.51% for climatic components alone, while the KNN model had the best accuracy of 90.71% for climatic elements alone. For socioeconomic and climatic factors taken together, the Random Forest model's accuracy was 87.82%; for climatic components alone, it was 88.83%. For cashew nuts, the accuracy of the KNN model was 78.38% for socioeconomic and climatic components combined and 99.81% for climatic factors alone, compared to 88.27% for socioeconomic and climatic elements combined and 86.58% for climatic factors alone for the Decision Tree model. For both socioeconomic and climatic components combined, the Random Forest model's accuracy was 98.50%, while for climatic factors alone, it was 98.75%. In conclusion, the Random Forest model outperformed the KNN and Decision Tree models across all crop and factor combinations. Our findings indicate that machine learning algorithms can be used to forecast crop yields with reasonable accuracy when socioeconomic and meteorological variables are combined

ACKNOWLEDGEMENTS

I want to sincerely thank Prof. Prasad for his consistent advice, inspiration, and support throughout my research process. His extensive knowledge, skill, and insights have greatly influenced my studies and professional development. I am appreciative of his guidance and for providing me with ongoing inspiration, challenges, and motivation.

Moreover, I want to sincerely appreciate my departmental colleagues for their unrelenting support, suggestions, and words of encouragement. I am appreciative of the chances they have given me to develop professionally and make a contribution to the scientific community.

My genuine gratitude also goes out to my family and friends for their strong support, love, and tolerance during my academic endeavours. Their encouragement has been instrumental in helping me get through the difficulties and to enjoying the experience.

Finally, I express my gratitude to God for his blessings, direction, and grace in enabling this accomplishment.

DEDICATION PAGE

I write my dissertation as a dedication to my friends and family. Special thanks go out to my devoted parents and sisters, whose words of support and push for persistence continue to reverberate in my ears. I also dedicate this dissertation to all of my close relatives and my church family for their help and encouragement during the writing process.

LIST OF ABBREVIATIONS AND TERMS

AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolutional neural networks
LSTM	Long Short-Term Memory
DL	Deep Learning
DNN	Deep Neural Networks
DT	Decision Tree
GDP	Gross Domestic Product
KNN	K-Nearest Neighbor
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
PPP	Purchasing Power Parity
RMSE	Root Mean Squared Error
SVM	Support Vector Machine

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER ONE

INTRODUCTION

A significant aspect of the Nigeria economy depends on agriculture. Interestingly, an estimate of 80% of cropland globally is used for rainfed agriculture, which produces strong yields when the weather is good for the crops.[1].

Nigerian agriculture, in particular that of the Lower River Benue Basin, has a number of difficulties that endanger its continued development. Some of these issues are brought on by socioeconomic dynamics that are concerned with access to land, farming practices, credit availability, poor processing and storage infrastructure, farm size, and input availability.[2]

The bulk of Nigerian households depend heavily on agriculture, which also contributes significantly to the country's economy [3]. The agricultural sector in Nigeria has significant economic benefits including the provision of food, GDP contribution, employment opportunities, provision of raw resources for agro-related industries and generation of foreign currency (the primary source of foreign exchange earnings up to the early 1970s was agricultural exports).

With a total size of 923768 square kilometers and an estimated population of 180 million, Nigeria is one of the biggest nations in Africa. It is located on the western coast of Africa, entirely within the tropics along the Gulf of Guinea. Due to Nigeria's highly diverse agro-ecological environment, a large variety of agricultural goods can be produced. As a result, one of the most significant economic sectors is agriculture.

With nearly 70% of the labor force employed, agriculture is by far the most significant sector of Nigeria's economy in terms of employment. Agricultural holdings are typically modest and dispersed; farming is frequently of the subsistence kind, distinguished by basic equipment and movable cultivation. It is estimated that 80% of all farm holdings are owned by smallholders, the majority of whom are subsistence farmers. And it is worthy to note that 80% of the food produced are by these small farms. Thus, the industry is particularly significant in terms of the number of jobs it creates, the GDP it contributes to, and the amount of money it makes through exports. 32 million hectares of Nigeria's 79 million hectares of arable land are under cultivation.

The two primary kinds of agricultural products produced in the country are food crops produced for both domestic use and exports. Some of the most important crops include beans,

sesame, cashews, cassava, cocoa beans, groundnuts, gum arabic, kola nuts, melon, millet, palm kernels, palm oil, plantains, rice, rubber, sorghum, soybeans, and yams.

The majority of agricultural productivity is fed by rainfall. Production of both crops and livestock continues to be below potential. Agriculture is not growing at the 10% rate required for achieving food security and reducing poverty. Shortfalls are caused, among other things, by poor availability to and low uptake of high-quality seeds, low fertiliser usage, and generally ineffective production practices. These factors can be grouped into socioeconomic and climatic. Climate conditions and the agricultural sector in Nigeria are directly correlated [4]. Most farmers rely on past down agricultural practices that are hinged heavily on weather conditions.

On the aspects on how climate affects farming practices, for instance, crop growth and soil erosion are impacted by rainfall intensity. The heating of the soil, plants, and their metabolic processes are also impacted by temperature. Although all crops are impacted by climate change, each crop is affected to varying degrees. One obvious example is the impact of temperature on cocoa, which grows best between the ranges of 18 to 32 °F. Temperature acts as a catalyst for biochemical processes such as photosynthesis and cellular respiration, and even little temperature variations can have an impact on crop productivity [5]. Also according to [6], monthly rainfall has a huge effect on agricultural production overall and is strongly correlated with growing season temperatures, crop maturity, and crop yield in crops like cowpea, maize, and rice.

Farmers require precise knowledge of the various seasons and how they impact crop productivity. Thus, getting accurate agricultural output predictions for each season is a problem that needs to be solved. To better optimize processes and take advantage of climatic conditions, industry stakeholders and policy officials must be able to predict yield. As crop selection is an important part of agricultural planning, this will help farmers take the necessary actions to increase output.

Accurate crop prediction is feasible by utilizing early prediction by data collection of previous farmers' experiences and environmental conditions such as rainfall data, humidity, sun radiation, and temperature and applying machine learning techniques [7]. Also, farmers may accurately predict crop production and expected profit by using machine learning to predict the crop yield and easily schedule and plan crop propagation, such as alternative crops to be planted or not under different climatic conditions, with the help of precise weather-

based crop forecast. Farmers can reduce future losses by being alerted beforehand, which will ensure company loss prevention and promote economic growth. [8].

1.1 Problem Statement

The majority of climate change research explains agricultural output in terms of Climate-related and biophysical elements like temperature, rainfall, and soil. Most research globally have focused on multiple machine learning algorithms; crop yield may be correctly predicted depending on climatic factors. But there exists a necessity to investigate the way socioeconomic and climatic factors interact to affect yield given both the agriculture industry's ongoing extensive economic reforms, investment and a fast-changing biophysical environment.

This research presents a machine learning algorithm that incorporates socioeconomic variables, such as GDP (Gross Domestic Product) power purchase parity, inflation, fertilizer use, and land use that have been found to affect agricultural yield in various ways. Researches have displayed that higher GDP per capita and power purchase parity are associated with higher agricultural yield, as they enable farmers to invest in better technologies, inputs, and infrastructure. On the other hand, high inflation rates can lead to higher input costs, which can reduce yield.

Research by [9] on the southern Canadian prairies' canola crop yields appear to be somewhat impacted by the usage of digital technologies. About 38% of the canola yield that were found in the study site in the Canadian prairies were explained using a combination of temperature and precipitation over the course of the three-month season of canola growth and four socioeconomic factors. The findings serve as a starting point for recommending to agricultural producers and policy makers which socioeconomic factors have the greatest impact on canola crop output. This research brings to the fore that crop yield prediction is not solely based on climatic factors.

Also, fertilizer use has been found to have a positive impact on yield, as it provides essential nutrients for crop growth. However, excessive use of fertilizers can lead to environmental problems such as soil degradation and pollution. Land use also affects yield, as factors such as soil quality, water availability, and topography can vary depending on the type of land use.

Socio-economic factors play a significant role in determining agricultural yield, and policymakers should take them into account when designing policies to promote sustainable agriculture and food security.

1.2 Aims and Objectives

1.2.1 Aim

This thesis aims to evaluate the performance of three machine-learning model on climatic and socioeconomic data for crop yield prediction

1.2.2 Objective

The research objective for crop production prediction based on climatic and socioeconomic factors using machine learning models include:

- To determine how accurate and reliable machine learning model can predict crop yields based on climatic and economic factors, with the aim of improving agricultural decision-making and supporting food security.
- Explore different machine learning techniques including random forest, knn and decision tree model to identify the performance of these model.
- The ultimate goal is to provide farmers, policy-makers, and other stakeholders with actionable insights that can help to optimize crop production and reduce the immense effects of climate change on the agricultural sector.
- Determine the statistical significance and compare the performance of crop yield model using climatic factors versus the model using climatic and socioeconomic factors

1.3 Research Questions

The research questions include:

1. What is the performance of the machine learning models used for predicting crop yield and comparing their performances with climatic data alone?
2. How can the insights gained from this study inform policymakers and farmers about the most effective strategies for improving crop yield?

3. Can the developed machine learning model be applied to other regions with different climate and socioeconomic conditions to predict crop yield accurately?
4. What are the limitations and challenges associated with using these machine learning methodologies for prediction of crop yield by combining socioeconomic and climatic data?
5. How can the findings of this research contribute to the existing body of literature on crop yield prediction and sustainable agriculture?

1.4 Scope of Study

The prediction of Crop yield is an important task for ensuring food security and sustainable agricultural practices. The yield of crops can be influenced by various factors such as climate variability, soil quality, and socioeconomic factors. This thesis aims to predict crop yield based on both climate variability and socioeconomic factors using machine learning techniques.

The scope of this research is limited to the following:

1. The thesis will focus on evaluate the performance of three machine-learning model on climatic and socioeconomic data for crop yield prediction
2. The research will consider hyper parameter tuning and cross validation and carry out training and testing of the three models

1.5 Significance of the Study

The combination of socio-economic data and climatic data can be useful in predicting crop yield in ways such as:

1. While Socio-economic data can provide information on factors such as economic conditions, land use, which can affect crop production, Climatic data, on the other hand, can provide information on factors such as temperature, rainfall, and soil moisture, which are essential for crop growth.
2. By combining these two types of data, researchers can gain critical insights into how machine learning models can predict crop yield more accurately and help farmers make informed decisions on crop management. For example, models that incorporate

socio-economic data can provide insights into the impact of policies such as subsidies or trade agreements on crop production.

3. The combination of socio-economic and climatic data can provide a more comprehensive and integrated approach to crop yield prediction and help address food security challenges.

1.6 Expected Results and Deliverables

1.6.1 Expected Results

The expected outcomes of this thesis are as follows:

1. Determining the machine learning model that can accurately forecast crop yield based on climate variability and socioeconomic factors.
2. Comparison of the models using the combined factors, with the models using climatic factors only
3. Gain valuable insights into how climate variability and socioeconomic factors interact to affect crop yield.

The thesis will contribute to the existing literature on crop yield prediction and help inform policymakers and farmers about the most effective strategies for improving crop yield. The results of the machine learning models can be used to determine its performance in predicting crop yield in different regions and can aid in the development of sustainable agricultural practices.

1.6.2 Deliverables

A journal paper will be published at the end of this research for evaluating the performance of machine learning models on crop yield prediction using climatic and socioeconomic factors

1.7 Research Outline

The research outline is as follows:

1. **Chapter one** contains an introduction to Nigeria's agricultural sector and weighs in on the factors such as climatic and socioeconomic affecting crop yield. It highlights a

problem statement, aims & objectives, the scope of the research, the significance of the research, expected results and deliverables, and then the thesis outline.

2. **Chapter two** contains the related work of research on socioeconomic factors affecting crop yield and machine learning models in evaluation crop yield based on climatic data.
3. **Chapter three** contains a discussion of our proposed model and the method we used to meet the research objectives.
4. **Chapter four** contains the performance evaluation of the models. First, we will compare the three models combining socioeconomic and climatic data and then compare them with results using climatic data only.
5. **Chapter five** contains the summary, conclusion of our research and possible open research directions.

CHAPTER TWO

RELATED WORKS

2.1 Socio-economic Factors Affecting Crop Yield

For sub-Saharan African nations like Nigeria, research on weather-based crop production prediction using a machine learning technique is not common, and none of the studies have addressed adding socioeconomic aspects. This study introduces a machine learning weather-based system that forecasts agricultural yield using socioeconomic and climatic variables.

Farmers around the nation have been exposed to a variety of climate change adaptation techniques. The options for adaptation, however, place an undue emphasis on technical knowledge and capabilities and neglect to take into account important social aspects like culture, values, and beliefs that affect how effectively new technologies, capacities, and skills for adaptation are adopted. This study [10] examined how socioeconomic characteristics in the villages of Pwalugu and Balungu of Ghana's Upper East Region affect farmers' individual processes in adapting to climate change. In this study, the communities were chosen using the purposive selection technique, and 100 respondents were chosen at random from the study communities. Focus groups, key informant interviews, and surveys were used to collect information from respondents. The data in this study were subjected to a thorough statistical analysis, and the obtained results were shown in figures and tables form. The research emphasizes the institutional and legal framework that must be used to in Northern Ghana's rural communities to the effects of climate change. Additionally, this suggests that the government and other relevant parties work with financial institutions to make sure that funds are easily accessible to farmers so they can foster adaption to climate change in an effective way. Farmers should also receive training or participate in workshops to improve their capacity for developing and putting into action effective climate change strategies. The integrated methodological approach was adopted in this study, which coupled appropriate quantitative procedures with qualitative ones. The approach guarantees the validity (the ways in which measurements are accurate) and reliability (the degrees to which results are consistent across time) of the research. Key informant interviews, alongside household

questionnaire surveys, and focus-group styled discussions were utilized as a combination of participatory approaches, giving locals the chance to contribute by offering their collective experiences and depth of knowledge to define potential solutions to the issue at hand. Several approaches are effective at minimizing the shortcomings of a single method. The research's design was based on a cross-sectional study. In a certain population, measurement or determination of variables were at the same time. This technique made it possible to evaluate a population's customs, attitudes, information, and beliefs in respect to a specific event or phenomena. According to the study's findings, farming was the main occupation in the two communities, and men predominated. Some activities that respondents engaged in to make a living included fishing, raising animals or poultry, producing firewood or charcoal, hunting, and driving. Regarding institutional frameworks, the principles guiding decision-making in the two communities were that bush burning and tree felling were avoided. The reasons these norms were followed in the research area were fear of punishment, some of the grasses being grazed by animals, trees causing rainfall, and rewards and incentives that respondents receive from trees. The socioeconomic elements in the research region were characterized as land access, gender dynamics, and finances. The difficulties involved in acquiring land in the communities included high demands from landowners, last-minute changes of heart by landowners, a lack of productive land, a lack of funding to acquire land, tenant behavior, the number of acres necessary, and properties remote from water bodies. The capacity of respondents to obtain fertile property, land that was near water, and any number of acres of their choosing was determined by their access to financing. Yet, gender limited women's ability to adapt to climate change. Just because they are viewed as immigrants and lack knowledge of the local history, women were also not permitted to own land or other types of property, such as animals.

The study [11] was conducted to evaluate the socioeconomic traits of groundnut farmers, ascertain the profitability level of groundnut production, the efficiency of resource use, as well as to identify issues faced in groundnut production in the study area of Sabon-gari local government area due to this significant gap. By the sale of seed, cakes, oil, and haulms, groundnut, an important oil seed crop, generates substantial amounts of revenue. The diets of rural people often include a lot of groundnuts. The average production of groundnut pods from farmer's fields is only about 800 kg per ha, which is less than one-third of the maximum yield of 3000 kg per ha. 79 farmers who produce groundnuts were chosen at random from among the many farms spread around the local government region. Using both primary and

secondary sources, data were gathered. The gross margin and cost-benefit analysis were performed to assess the profitability of groundnut production. The research findings indicate that seasoned farmers are less likely to produce groundnuts and that the majority of groundnut farmers are active in other types of industries. Poor use of the inputs is caused by their cost, availability, and lack of technical knowledge of the needs. Except for insecticide, which is underutilized, fertilizer, seeds, labor and herbicides are all overused. Lack of funding and extension services are two issues that groundnut production in the study runs into. Around 78% of the groundnut issue in the research area was due to these two issues. Therefore, it is advised that government and research organizations in strengthening their extension services to provide farmers with better technologies. Farmers ought to look for loans through cooperatives, banks, and other affordable sources is also recommended. The loan application process should also be simplified simple to make it easier for farmers to acquire loans and increase peanut production.

The study [9] looked into the socioeconomic aspects of agricultural infrastructure that affect its accessibility, availability, and satisfaction for smallholder farmers. Using cross-sectional data from the South African North West Province a total of 150 smallholder farmers were chosen using stratified sampling, which divided the farmers into those who had access to agricultural infrastructure and those who did not. STATA 14.0 was used to code, collect, and analyze the data. Descriptive analyses and Tobit Regression Models were employed in the analysis. According to the Tobit Regression Model's findings, household members' assistance in farming enterprises, farm ownership, farm acquisition, farmer occupation, membership in farmer organizations, sources of labor and farming experience, and agricultural production inputs all played important roles in determining the availability of agricultural infrastructure. The following factors were crucial in determining the accessibility of agricultural infrastructure: involvement in non-farming activities, interaction with extension services, farm ownership, farmer occupation, membership in farmer organizations, labor availability, farming experience, and land tenure. The following variables, among others, played a crucial role in determining how satisfied farmers were with agricultural infrastructure: farm ownership, membership in farmer organizations, farmer age, education level, marital status, and gender; household members' assistance in the farming enterprise; farmer receiving government agricultural support. The analysis' findings were utilized to fill in knowledge gaps regarding how North West Province smallholder farmers' productivity and revenue from

agriculture are affected by agricultural infrastructure, availability, accessibility, and satisfaction.

This study's [12] objective was to evaluate the effect of socioeconomic variables on maize yield between 2016 to 2017 in Tanzania's southern and northern maize production zones. This survey results were from the Bringing Maize Agronomy to Scale in Africa (TAMASA) initiative, which covered 8 districts in Tanzania. The study's regions' average maize yields in 2016 and 2017 varied greatly, according to the survey. The years were very different. Variations in plant density at harvest time caused the biggest portion (13%) of the difference in maize output among 8 distinct regions of Tanzania.

Although its production is still limited, pumpkin is an indigenous produce with enormous potential to help Kenyan households with nutrition, food security, and income. The crop has not been promoted as a profitable business and has received little research interest. According to the literature, no studies or records have been made of the effects of socioeconomic factors and farming limitations on smallholder farmers' production, consumption, and selling of pumpkins in Eastern and Central Kenya. In order to better guide the creation of relevant policy interventions for increased pumpkin production, consumption, and marketing, this research was conducted to evaluate these drivers and limits. Eight significant pumpkin-growing Sub-Counties in the semi-arid regions of Eastern Kenya and the medium-altitude regions of Central Kenya were the sites of the study [4]. The study's goals were to: (a) determine how socioeconomic and demographic factors affect pumpkin production; (b) examine how smallholder farmers use pumpkin products and adhere to sociocultural customs when eating pumpkin; (c) identify market characteristics that affect pumpkin marketing; and (d) pinpoint and analyze the main barriers to smallholder farmers' ability to produce and market pumpkin in the Eastern and Central Kenya regions. Using structured questionnaires, a household survey of 260 pumpkin-growing households and a market survey of 172 main dealers were carried out. With the use of SPSS and Stata computer software, the acquired data were examined using descriptive statistics, multiple regression, and Tobit model analysis. According to the study, smallholder farmers in Eastern and Central Kenya produced less pumpkin per hectare than the national average of 20 tons. Smallholder pumpkin production in Eastern and Central Kenya was statistically significant and positively influenced by participation in off-farm activities, household size, on-farm income, farm area under pumpkins, and household head's age and education level. The majority of households mostly used seeds for sowing whereas pumpkin fruits, leaves, and seeds were primarily used

as food. Among farm homes in Eastern and Central Kenya, household size and proximity to the market had a statistically significant negative impact on the proportion of marketed pumpkins. In Eastern Kenya, belonging to a farmer's organization was important, whereas in Central Kenya, market price and the gender of the family head were important. These elements influenced the percentage of marketed pumpkin among farm households favorably. Market involvement by pumpkin vendors in Eastern and Central Kenya was statistically significant and positively influenced by market pricing, membership in marketing organizations, frequency of sales, and distance to market. Pests, diseases, and a lack of rainfall were the key production restrictions for pumpkins, whereas bad market prices, broker exploitation, post-harvest losses, a lack of market intelligence, low consumer awareness, and low demand were the top marketing constraints. The suggested policy interventions involve educating farmers, promoting pumpkin production, improving access to information and physical marketplaces, grouping farmers for marketing purposes, enhancing market infrastructure, and forming associations or clubs for pumpkin traders.

This study [13] analyzed the effect of farmers' purchasing power on Nigeria's agriculture industry and considered the level of the nation's current food supply. The findings of this study demonstrate that, despite the fact that Nigeria's food production is currently reducing poverty to a greater level, poverty has persisted in the nation because farmers are ignored and given few possibilities. Additionally, it was noted that the absence of support from the authorities deters aspiring farmers and agriculturalists from pursuing their careers. It reinforced the necessity for farmers to have access to low-interest financing so they can expand their farms.

Also, this author [14] conducted a crop production analysis and discovered that the main causes of the performance of yam production in the research area were the farmers' age, education, farming experience, farm distance, and income level. These factors all had positive coefficients and were statistically significant as socioeconomic factors.

A field experiment by Gopal et al [15] was conducted in Gujarat from 2009 to 2011 to evaluate the effects of different treatments on pearl millet growth and soil properties. Four treatments were tested: control (T1), farmyard manure (FYM) at 5 tonnes/ha/year (T2), FYM at 5 tonnes/ha/year + N:P:K @100:60:40 every year (T3), and FYM at 10 tonnes/ha/year + N:P:K @100:60:40 every year (T4). Results showed that T4 had the highest plant height, biomass, and yield. T2 and T3 had intermediate growth parameters, while T1 had the lowest.

T4 also had the lowest runoff and soil loss, and the highest soil organic carbon (SOC) content. SOC was highest in water stable aggregates (WSA) of size >0.5 mm and in the top 15 cm of soil. Overall, application of FYM at 10 tonnes/ha/year + N:P:K @100:60:40 resulted in better crop growth, higher yield, lower runoff and soil loss, and higher SOC.

As yields improve and environmental costs rise as a result of complex interplay between social, economic, and ecological issues, ensuring food security while minimizing environmental costs is agriculture's fundamental concern. In this study [16], the author examined to determine the outcomes of socioeconomic factors on the effectiveness of grain production between China and Ethiopia. A number of measures on land reforms, from community systems to tax cancellation and subsidies, were put in place to set the economic transformation and increase grain yields in rural China. Ethiopia had also undergone several forms of land reform, from the landlord and peasant system to land as the common property of Ethiopia's nations, nationalities, and peoples. In the 1980s, the two countries experienced nearly identical growth in terms of their gross domestic products per capita, which is a measure of the average level of living for citizens in a nation. Later, however, there was a major difference between the two nations. It suggested that in China, policies that minimize the fertilizer inputs were largely advised to reduce the environmental costs associated with higher agricultural inputs for sustainable agriculture growth. In Ethiopia, it suggested that infrastructure development that are better and meet socioeconomic demands needed to be given priority in order to meet food security and increase agricultural efficiency.

Table 1. Selected Socioeconomic factors affecting crop yield in Nigeria

S/N	Item	Definition	Author
1	Purchasing Power Parity	When comparing the fiscal output and the standard of living of people, purchasing power parity (PPP) is a common measuring method used for macroeconomic assessment. PPP is a financial concept that uses a "basket of goods" approach to compare the currencies of various countries. Instead, using exchange rates from the world markets, which may distort the true disparities in per capita income, it additionally considers the relative costs of local goods, services, and inflation rates. The cost of living is perhaps a more popular name for the idea of purchasing power parity.	[13]
2	Inflation	Price increases, sometimes known as inflation, are essentially the progressive decline in purchasing power. The pace of reduction in	[17]

		purchasing power can be approximated by the mean price increase of a sample of goods and services over time. Because of the price increase, which is commonly expressed as a percentage, one unit of money actually buys less. Inflation is comparable to deflation, which occurs when prices decline and purchasing power increases.	
3	Fertilizer Use	Fertilizers are extra materials that are given to the crops to boost productivity. Farmers utilize these on a regular basis to boost crop productivity. These fertilizers contain nitrogen, potassium, and phosphorus, three essential nutrients that plants require. They also improve the soil's fertility and water-holding capacity.	[18]
4	Land Use	One of the key elements affecting the amount of agricultural land is the labor required. Previous research has demonstrated that the need for labor increases with farm size.	

2.2 Climatic Factors Affecting Crop Yield

The prediction of crop yield uses a number of models, including Support vector machines, decision trees, artificial neural networks, naive bayes, linear regression, and logistic regression, among others. There is not a lot of research on predicting crop yields depending on weather in west African countries. Hence, a machine learning weather-based system that forecasts crop yield based on climate variability is presented in this study.

Thomas et al [19] conducted a comprehensive analysis of machine learning techniques for agricultural yield prediction. According to the data, the most frequently used characteristics are temperature, rainfall, and soil type, and the most frequently used methods are artificial neural networks, random forest, linear regression, and gradient boosting tree. Deep learning algorithms like Convolutional neural networks (CNN), Long Short-Term Memory (LSTM), and Deep Neural Networks (DNN) were all commonly used in studies that did so. The study also made clear that the choice of features is determined by the available datasets and the study's objectives.

Rice, maize, cassava, seed cotton, yams, and bananas were the six crops that the author Cedric et al [20] utilized for his research. The crop k-Nearest Neighbor (Ck-NN) model achieved the highest overall rating among the three models, according to the data. Whereas the R2 of the Crop Decision Tree (CDT) and "Crop Multivariate Logistic Regression (CMRL) models were 94.65% and 83.80%, respectively, on test data, it had an R2 score of 95.03% and an MAE of 0.160 kg/ha. They also looked at how well each model performed

when applied to crops, and the results revealed that the Ck-NN model performed best and the CMLR model performed worst

Mengjia et al. [21] proposed SSTNN (Spatial-Spectral-Temporal Neural Network), a novel deep learning architecture that combined 3D convolutional and recurrent neural networks for agricultural productivity prediction. The results were compared to both current deep learning methods and popular machine learning strategies. The proposed SSTNN was compared to competing deep learning techniques in order to further evaluate the superiority of the offered strategy (CNN and LSTM). The comparisons showed that the SSTNN beat the CNN and LSTM in the prediction of both corn and winter wheat on all metrics (i.e., RMSE, R2, and MAPE). For the prediction of winter wheat, SSTNN outperformed CNN in terms of RMSE and R2 by 20.2% and 12.2%, respectively, and LSTM in terms of RMSE and MAPE by 26.3% and 26.5%, respectively. In comparison to CNN, SSTNN's prediction of corn yield had a lower RMSE of 0.13 and a higher R2 of 0.10. When compared to the LSTM, the MAPE of the SSTNN was 28% lower. Also, the research's findings demonstrate that there was a lack of spatial, spectral, or temporal information that affected both machine learning and other deep learning techniques. The reality demonstrates that merging spatial-spectral and temporal data can significantly increase crop yield prediction accuracy. Furthermore, the study found that neither the suggested method nor the competing approaches were very effective in predicting corn yields. This is probably because corn had a shorter growth season than wheat, which means there were fewer temporal data available for prediction.

Saeed et al [22] painstakingly created deep neural networks that could learn nonlinear and complex correlations between genes, environmental factors, and their interactions from historical data in order to forecast the yields of novel hybrids planted in unknown sites with known weather conditions. The model's performance was discovered to be rather sensitive to how well the weather was predicted, which indicated the significance of weather prediction methods. Although many machine learning techniques share the black box trait, it is a significant drawback of the suggested model. They used the backpropagation method to do feature selection that was on the trained DNN model to lessen the model's black box nature. With a validation dataset using expected meteorological data and a root-mean-square-error (RMSE) of 12% of the average yield and 50% of the standard deviation, it was discovered that this model had a higher level of prediction accuracy. The RMSE was seen to decrease to 11% of the average yield and 46% of the standard deviation with ideal meteorological data. Also, they used feature selection based on the trained DNN model, which was successful in

reducing the input space's size without noticeably lowering prediction accuracy. According to the computational results, this model performed much better than other well-liked techniques like Lasso, shallow neural networks (SNN), and regression tree (RT).

Alexandros et al [23] also constructed deep learning-based models to examine how the base algorithms fair in terms of numerous performance criteria. XGBoost as a single model, XGBoost with scaling, XGBoost paired with scaling and feature selection methods, and hybrids Deep neural networks (DNN) powered by CNN-XGBoost Convolutional Neural Networks (CNN) and CNN-Recurrent Neural Networks (RNN),and CNN-Long Short-Term Memory were the algorithms evaluated in the study (LSTM). A public soybean dataset with 395 attributes, including weather and soil conditions, and 25,345 samples was used in the study's experiments. The hybrid CNN-DNN model performed better than other models, according to the data, with an RMSE of 0.266, an MSE of 0.071, and an MAE of 0.199. It also came to the conclusion that the CNN-RNN model can perform with an R2 between 85.4 and 87.09% and an RMSE between 4.15 and 4.91. According to the study, the CNN-RNN model performed somewhat worse (R2 = 77%, RMSE = 0.350), whereas two of the other recommended models produced results that were almost same but required less processing power, were less sophisticated, and used more specialized methods. The model's predictions fit with an R2 of 0.87. The XGBoost model, which executed faster than the other DL-based models and produced a second-best result with R2 = 83% and RMSE = 0.299. A limitation of the research is the use of the dataset for the soybean crop in the Corn Belt of the United States. because other datasets could yield somewhat different results when the suggested models are used.

There is still potential for additional research to make the models more understandable, despite the fact that the research did feature selection and set a variance threshold to describe which features should be employed. Further information about the features that were chosen and how they were chosen should be available. For predicting crop yield from sequential data of dates, a hybrid model like XGBoost that combines an attention mechanism and a DL algorithm like RNN or LSTM may perform better.

The majority of machine learning models now in use base their forecasts on NDVI data, which may be challenging to utilize because of clouds and their related shadows in collected photographs as well as the lack of trustworthy crop masks for broad areas, particularly in developing nations. In this study, the author demonstrated a deep learning model that could

forecast five distinct crops both before and during the growing season. [24]. Using crop calendars, readily available remote sensing data, and weather forecast information, our program delivers accurate production estimates.

This paper outlined the creation of an upgraded corn yield projection model for the Midwest of the United States (US). Using satellite pictures and meteorological data from the dominating expansion phase, they were able to evaluate six different artificial intelligence (AI) models. The paper defined the drought and heatwave by taking into account the traits of maize growth and chose the cases for sensitivity tests from a historical database in a bid to assess the effects of extreme weather events. The deep neural network (DNN) model's hyperparameters were precisely adjusted to provide the best configuration for accuracy growth [25] .

The author [26] conducted an experiment to predict wheat yield using canopy reflectance spectra. Red Normalized Difference Vegetation Index (RNDVI), Green Normalized Difference Vegetation Index (GNDVI) and Simple Ratio (SR) were used as spectral reflectance indices. Higher yield was observed in 0.8 and 1.0 irrigation levels than 0.4 and 0.6 irrigation levels. The model accounted for 79% variation in grain yield and 86% variation in biomass yield with slightly underestimated values.

A study by Ashis et al [27] was conducted to estimate the acreage of mango in West Godavari district of Andhra Pradesh, India, in 2017 using Sentinel 2 satellite data. Three classification techniques were used to prepare a land use and land cover map, and the Support Vector Machine with RBF kernel was found to be the most accurate, with an overall accuracy of 94.44% and a kappa coefficient of 0.9218. The estimated mango area was 9372.96 ha.

Also, Support vector regression, polynomial regression, and random forest were used to examine the data that had been obtained for irish potatoe and maize in Rwanda. Temperature and rainfall were utilized as forecasters. The models underwent testing and training with root mean square errors of 510.8 and 129.9 for maize and potatoes, respectively, and R2 values of 0.875 and 0.817 for the same agricultural datasets, the results show that Random Forest is the best model [1].

A summary of the various methods for carrying out the prediction of crop yield based on climatic factors are in Table 2: Comparative study of various models for crop yield prediction based on climatic factors

Table 2: Comparative study of various models for crop yield prediction based on climatic factors

Reference	Year	Machine Learning Models	Strength	Weakness	Accuracy (R^2)
[20]	2022	k-Nearest Neighbour	After storing all of the previous data, a new data point is categorized using the K-NN algorithm based on similarity. No time was spent training for classification or regression. The KNN algorithm does all its work during prediction and does not include an explicit training phase.	Does not function well with large dimensionality as this will make computing distance for each dimension more difficult. KNN is sensitive to dataset noise.	95.03%
		Decision Tree	The suggested crop decision tree enables the development of sophisticated algorithm that predict yield target variable or the establishment of a system based on many covariates. Missing values in the data have no impact on how the decision tree is constructed. Both data scaling and data normalization weren't necessary.	These findings point to an overfitting model within the DT model; hence the cross-validation method needs to be used.	94.65%
		Multivariate Logistic Regression MRL	In contrast to yam, the model gained more knowledge from cassava, cotton, rice, and banana. The fact that maize bears the lowest score compared to the other crops may indicate that the model has not learnt a lot on the maize data.	The results demonstrate that the MLR model executes more slowly than the decision tree and the k-NN, taking two times as long. This puts it at a disadvantage. Moreover, maize and cotton were forecasted at very low rates, indicating that these two crops had poor knowledge. The negative R2 score for	83.8%

				seed cotton indicates that there is less linear correlation between the prediction parameters and that crop's regression slope. This might be a contributing factor in the model's subpar performance.	
[28]	2021	Random Forest	Random forest can handle large datasets with high dimensionality, which is often the case for climate data. It is less prone to overfitting than other machine learning algorithms, as it uses multiple decision trees and averages their predictions.	The interpretability of the model may be limited, as it is based on multiple decision trees and the importance of each feature may not be easily understood.	87%
[21]	2021	Spatial-Spectral-Temporal Neural Network (SSTNN)	The acquisition of geographical, spectral, or temporal data is a challenge for machine learning as well as other deep learning techniques. The reality demonstrates that merging spatial-spectral and temporal data can significantly increase crop yield prediction accuracy.	The training phase of NNs is typically time-consuming because there are many parameters that need to be optimized. The suggested SSTNN requires more time to train—several hours—than most alternative algorithms.	Wheat - 83% Corn - 68%
[22]	2019	Deep Neural Network	DNN can capture the intrinsic nonlinearities in meteorological data and learn these nonlinearities from data without needing the nonlinear model to be set up before estimation.	The black box property of the suggested model, which is a characteristic of many machine learning techniques, is a significant drawback. Although the model reflects GxE interactions, it is complex in developing testable hypotheses that might offer biological insights	85.46%

				because of the model's complicated model structure.	
[23]	2022	CNN-DNN (with Robust Scaler and Select from Model)	With the right data processing techniques, architecture, and hyperparameter settings, it handled the huge dataset with ease. This model enabled them to accomplish their highest R2 performance and significantly lower RMSE.	Further research is still needed to improve the models' level of explanation. Further information about the features that were chosen and how they were chosen should be available.	87%
		CNN-XGBoost (with Robust Scaler)	It used CNN and the speed and efficiency of XGBoost to extract information from the data computation speed and capture the interdependence of the data.	Lacking accuracy, complexity increased when using one or more Deep Learning algorithms, making them challenging to manage and modify.	78%
		XGBoost (raw)	Computation efficiency and speed are both high-speed and accurate. Maybe as a result of its architecture's use of gradient-boosted decision trees, It was made to operate quickly and effectively in applications for both classification and regression.	low precision and speed of performance. It appears as though their capacity to manage time slows them slower.	80%
		XGBoost (with Robust Scaler)	The variation of the data and the various units should be taken into consideration while scaling. It was necessary to identify the outliers that affected the variance of the 395 various attributes in this situation since they had varied value ranges and measurement units. Because of this, the Robust Scaler approach outperformed the competition.	It displayed excellent speed but only average accuracy.	79%

		XGBoost (with Robust Scaler and Select from Model)	Possibility of boosting trees for improved model accuracy and speed. This model features a low RMSE, high computational speed, and high computational complexity.	Its level of intricacy is adequate.	83%
		CNN-LSTM (with Robust Scaler and Select from Model)	The XGBoost was utilized as an estimate in the feature selection process. The research also used CNN and the speed and effectiveness of XGBoost to find the dependencies in the data and deduce information from the data.	Complexity increases, but accuracy and speed decline	67%
		CNN-RNN (with Robust Scaler and Select from Model)	It can get a performance of R2 between 85.4 and 87.09%, along with an RMSE between 4.15 and 4.91.	It showed low precision and speed of performance. Their capacity to manage time appears to slow them down. less complexity, more specialized procedures, and fewer computational resources	77%

CHAPTER THREE

METHODOLOGY

This chapter discusses the methodology and the algorithms used in this research. It begins with the dataset description and sources and moves to discuss how the proposed model is built to achieve the research objectives. The methodology for this thesis will involve the following steps:

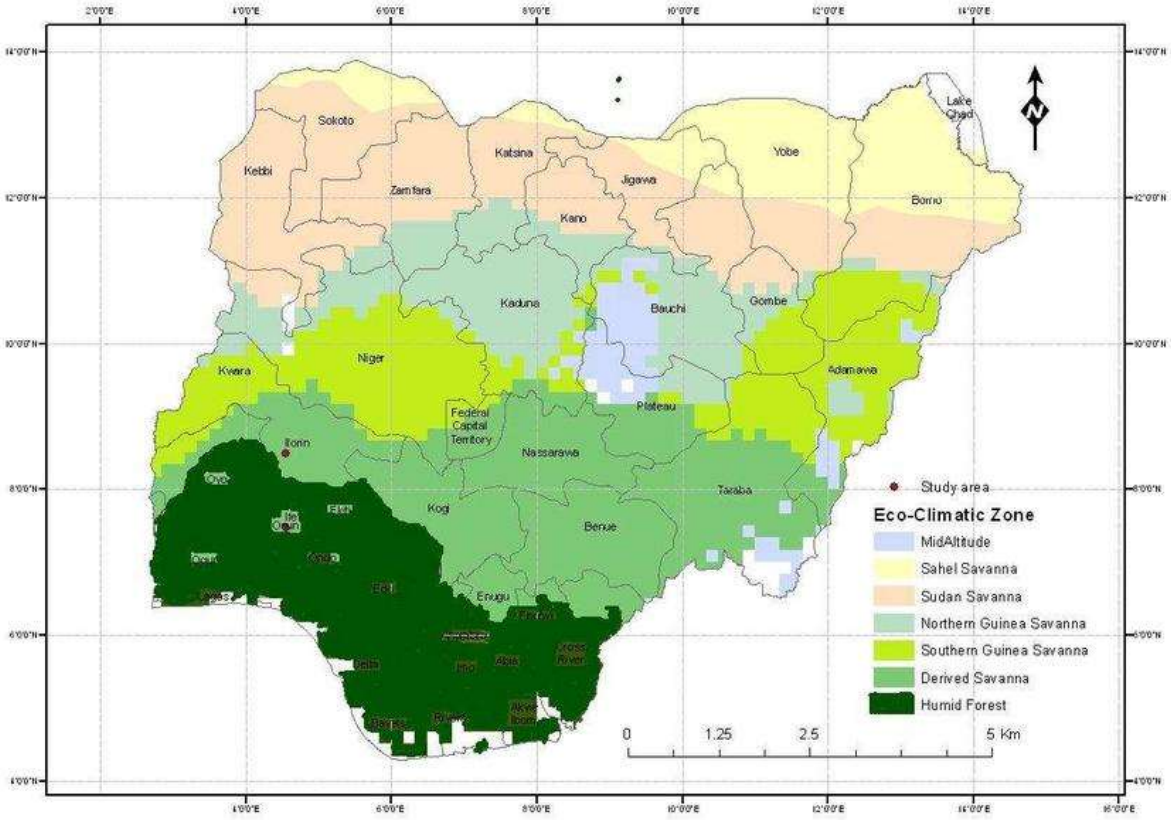
1. Data collection: Collect data on climate variability (e.g., temperature, rainfall, drought), soil quality (e.g., pH, nutrient levels), and socioeconomic factors (e.g., land use, access to technology) from publicly available sources and databases.
2. Data pre-processing: Pre-process the collected data to ensure it is clean and ready for analysis.
3. Machine learning model training and testing: Training and testing the machine learning models that predicts crop yield based on the both factors. The model will be trained using various regression techniques such as kNN, decision trees, and random forests.
4. Model evaluation: Evaluate the execution of the modesl using appropriate evaluation metrics such as mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R2).
5. Comparison with performance of the model using climatic factors only: Compare the performance of the models with the performance of these same models with climatic factors only.
6. Interpretation of results: Interpret the results obtained from the machine learning models to identify the best performing and see if there is a better prediction of yield of crops.
7. Statistical Significance Testing: Perform statistical significance test to determine the performance and statistical significance of the models

3.1 Study Site

Nigeria is a nation in West Africa that is surrounded by the Sahara Desert to the north and the Atlantic Ocean to the south. Between latitudes 4°N and 14°N and longitudes 4°E and 14°E are its respective locations. Because of this, the country experiences an incredibly broad range of climate variations throughout the year.

Nigeria has four distinct climate zones: monsoon climate in the Niger-Delta, warm semi-arid climate in other parts of the northeast, and tropical savannah in the middle belt and certain sections of the southwest. Nigeria's three primary ecological regions are semiarid regions in the north, savannah in the middle, and tropical rainforests in the south. The Eco climatic zones in Nigeria are shown in Fig 1 below.

Fig 1: Eco climatic zones in Nigeria [29]



The Federal Republic of Nigeria, also referred to as Nigeria, is located on the interior of the Gulf of Guinea on the west coast of Africa. Benin, Niger, Chad, and the Gulf of Guinea in the Atlantic Ocean make up its western, northern, eastern, and southern borders, respectively. The total area of Nigeria's land is 923,768 km2, including its 853 km of coastline. Nigeria is predominantly in the lowland, humid tropics, where it experiences high temperatures year-round, a somewhat moist coastland, and extremely desert northern regions. The high plateau

of Nigeria, which is located 300 to 900 meters above sea level, and the low plateau. and the lowlands, which typically aren't higher than 300 meters (Figure 1). The Eastern and North Eastern Highlands, the Western Uplands, and the North Central Plateau are a few of the high plateaus. All of western Nigeria's interior coastal lowlands, the Niger-Benue Trough, the Chad Basin, the lowlands and scarp lands of south-eastern Nigeria, the Sokoto Plains, and coastlands are regarded as lowlands. The low-lying Niger Delta, where the river Niger eventually drains into the sea, is defined by a sophisticated network of canals created by both man and nature.

In Nigeria, vegetation generally follows rainfall patterns and can be categorized as either tropical rainforest or savannah. Mangrove swamps, freshwater swamps, and high forests are some of the several types of forests. Guinea, Sudan, and Sahel are the savanna subtypes. The Guinea savannah, which spans the entirety of the nation, is the greatest vegetation belt there. The coastal strip, continually inundates the land with brackish water, mangrove vegetation can be found. Since 2012, Nigeria's economy has been the largest in Africa. It is a lower middle-income nation. 4 Nigeria's population is predicted to be 206.14 million in 2020, with a 2.5% annual population growth rate. Nigeria's population is anticipated to grow to 262.9 million in 2030 and 401.3 million in 2050, respectively.

By 2030 and 2050, respectively, it is anticipated that 60% and 70% portion of the population will reside in urban areas, up from the current 50%. 6 Gross Domestic Product (GDP) for the nation is expected to reach $432.29 billion in 2020, with an annual growth rate of 2.2% in 2019 and 01.8% in 2020. 7 Oil price volatility caused growth to reach a high of 8% in 2006 and a low of 1.5% in 2016, with the (GDP) growing at an average pace of 5.7% each year between 2006 and 2016. Despite the fact that Nigeria's economy has fared better recently than it did during earlier boom-bust oil price cycles, such as those that occurred in the late 1970s or the middle of the 1980s, oil prices still heavily influence the nation's growth pattern.

3.2 Climate Data

Global data on the climate's past, present, and future vulnerabilities and effects are available via the Climate Change Knowledge Portal (CCKP). The ERA5 (reanalysis) collection satellite provides several climatic variables from 1950 to 2020.

Relative humidity, precipitation, lowest temperature, maximum temperature, and mean temperature over the period 1990–2020 are among the climate variables taken into account for this study.

3.3 Socio economic data

The World Bank Indicators Databank was used to get the socioeconomic statistics. A collection of time series data on a range of topics is contained in the analysis and visualization tool known as Databank. The main World Bank collection of development indicators is called World Development Indicators (WDI), and it is compiled from officially recognized international sources. It comprises national, regional, and global estimations and provides the most recent and reliable data on world development.

In this study, socioeconomic parameters from 1990 to 2020 were taken into account, including age, educational attainment, purchasing power parity, and the area of agricultural land.

3.4 Agriculture Data

FAOSTAT database was used to get the agricultural data. The FAOSTAT database offers access to food and agricultural statistics for more than 245 countries and territories from 1961 to the most current year available. The dataset for Nigeria is sourced with the URL https://fenix.fao.org/faostat/internal/en/#country/159. These data sets span the years 1961 through 2020 and contain production/yield and harvested area. Because they are the main cash crops in Nigeria, this research has solely taken into account sesame seed, cocoa, and cashew.

3.5 Data Set Description

The dataset for the study on the prediction of sesame, cashew, and cocoa crops based on climatic and socioeconomic factors from 1990 to 2020 consists of two main categories of data: climatic data and socioeconomic data.

The climatic data includes information on relative humidity, temperature, and precipitation, which are key climatic factors that affect crop growth and yield. This data is collected on a annual basis for each crop type and region.

The socioeconomic data includes information on GDP PPP, inflation, fertilizer use, and land use, which are important factors that influence agricultural production and crop yield. This data is collected on a yearly basis, and covers the same time period as the climatic data.

The dataset covers three major cash crops: sesame, cashew, and cocoa, which are important sources of income for farmers and contribute significantly to the economy of the region. It is expected that the dataset will provide valuable knowledge into the relationships between climatic and socioeconomic factors and crop yield, which can help to inform agricultural decision-making and policy development.

3.6 Pseudo Code

Here's a pseudo code for implementing a random forest, decision tree and KNN model for crop prediction based on climatic and socioeconomic data:

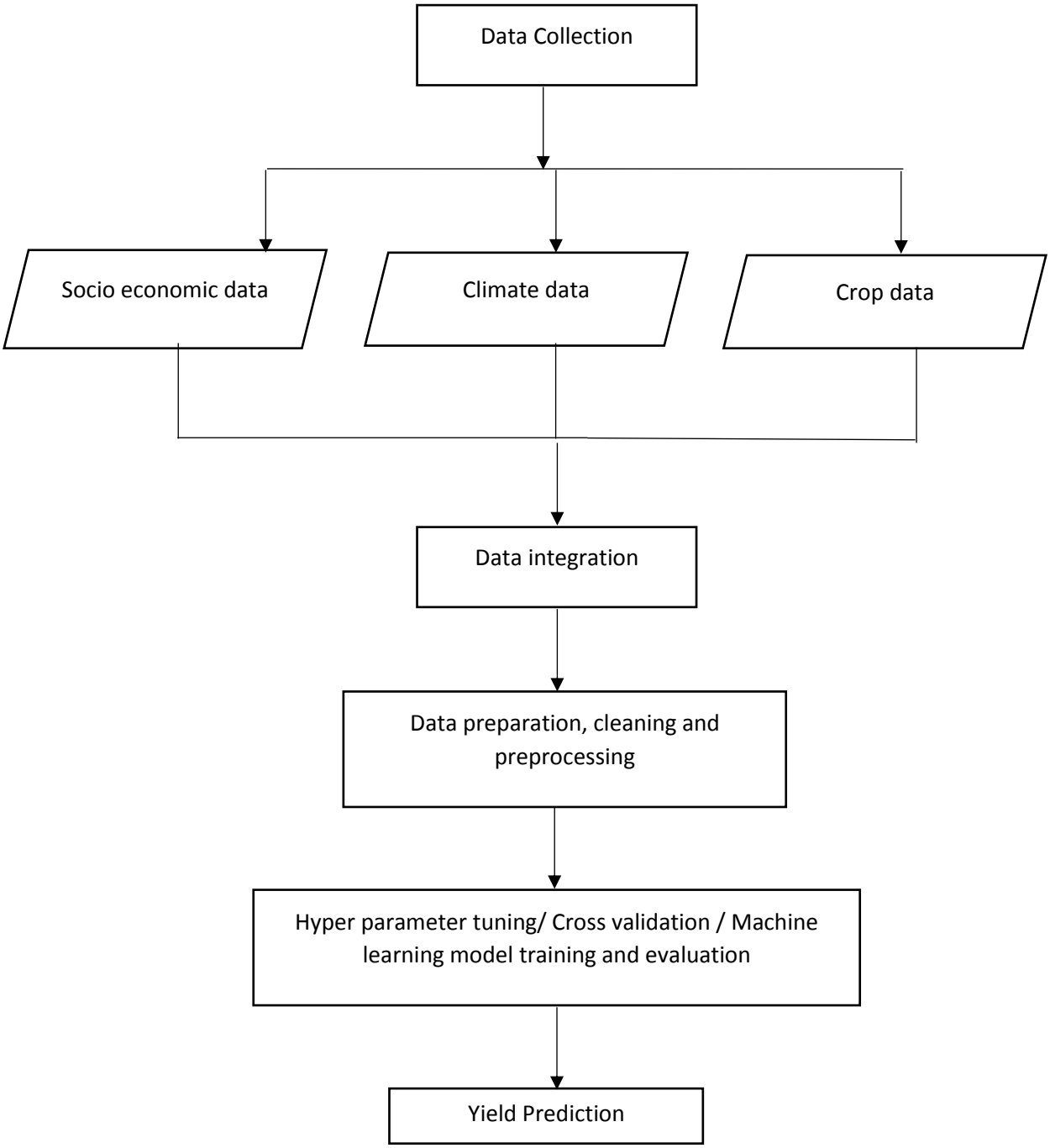
- Load the dataset of climatic and socioeconomic data for sesame, cashew, and cocoa crops from 1990 to 2020
- Preprocess the data by cleaning and transforming it into appropriate formats for machine learning
- Split the data into training and testing sets
- Define the random forest model with hyperparameters such as number of trees, maximum depth, and minimum sample leaf size
- Train the random forest, decision tree and KNN model on the training set
- Evaluate the model's performance on the testing set using accuracy metrics such as mean absolute error, root mean squared error, and R-squared value
- Use feature importance analysis to identify the most important climatic and socioeconomic factors for predicting crop yield
- Use the trained random forest, decision tree and KNN model to make predictions for future crop yields based on new climatic and socioeconomic data
- Visualize the results using appropriate charts and graphs to communicate the model's performance and predictions to stakeholders.

- Perform statistical significance test

The implementation of a random forest, decision tree and KNN model for crop prediction based on climatic and socioeconomic data involves several key steps, including data preprocessing, model definition, training and evaluation, and prediction. By following these steps and using appropriate evaluation metrics and visualizations, it is possible to develop an accurate and reliable machine learning model that can help to improve agricultural decision-making and support food security.

The flow chart is shown in Fig 2 below:

Fig 2: Flow diagram of the crop prediction model



3.7 Model Evaluation

Model evaluation is an important aspect of the machine learning process that involves assessing the performance of a model. The most common metrics used in model evaluation are accuracy, precision. The accuracy metric measures the correctly classified instances proportion, while precision measures the proportion of true positives among all positives.

In evaluating a model, it is important to use different metrics and consider the problem's nature. In this study, we used the accuracy score to evaluate the model.

3.7.1 Formula for Evaluation

Some key formulas that were used in the process of building our models for crop prediction based on climatic and socioeconomic data include:

3.7.1.1 **Mean Absolute Error (MAE):** A measure of the average absolute difference between the predicted and actual values of the crop yield. The formula for MAE is shown in equation 1 below:

$$\text{MAE} = \frac{1}{n} * \sum \left| y_i - y_{hat_i} \right| \quad \text{..... equation 1}$$

where:

- y_i being the actual value of the crop yield for the i-th sample
- y_{hati} being the predicted value of the crop yield for the i-th sample
- n being the total number of samples

3.7.1.2 **Root Mean Squared Error (RMSE):** A measure of the square root of the average squared difference between the predicted and actual values of the crop yield. The formula for RSME is shown in equation 2 below:

$$\text{RMSE} = \sqrt{[(\frac{1}{n}) * \sum \left(y_i - y_{hat_i} \right)^2} \quad \text{..... equation 2}$$

where:

- y_i being the actual value of the crop yield for the i-th sample
- y_{hati} being the predicted value of the crop yield for the i-th sample
- n: being the total number of samples

3.7.1.3 **R-squared (R^2) value:** This measures how well the model fits the data, indicating the variance proportion in the dependent variable (crop yield) that is explained by

the independent variables (climatic and socioeconomic factors). The formula for R^2 is shown in equation 3 below:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \text{..... equation 3}$$

where:

- SS_{res} being the sum of squares of residuals (the difference between the actual and predicted values of crop yield)
- SS_{tot} being the total sum of squares (the difference between the actual crop yield and its mean value)

3.7.1.4 **Mean Absolute Percentage Error (MAPE):** It is a commonly used metric to measure the accuracy of a forecasting model or prediction. MAPE measures the average absolute percentage difference between the actual values and the predicted values. MAPE is expressed as a percentage, and lower values indicate better accuracy. A MAPE value of 0% would indicate a perfect prediction model, while a high MAPE value indicates a larger discrepancy between the actual and predicted values. The formula for MAPE is shown in equation 4 below:

$$\text{MAPE} = \frac{1}{N} * \sum \frac{|(\text{Actual Yield} - \text{Predicted Yield})|}{\text{Actual Yield}} * 100 \text{..... equation 4}$$

Where:

- N: number of observations or crop yield predictions
- Actual Yield: the actual crop yield value for a particular observation
- Predicted Yield: the predicted crop yield value for a particular observation

3.8 Hyperparameter Tuning

Machine learning algorithms have various parameters that need to be set before training. These parameters are often referred to as hyperparameters, and their values can significantly impact a model's performance. Examples of hyperparameters include the decision tree maximum depth, the number of neighbors in KNN, and the number of trees in Random Forest.

Hyperparameter tuning is the process of selecting the best hyperparameters for a particular model. This is often done through a search process that involves trying different hyperparameter values and evaluating the model's performance for each. There are various

approaches to hyperparameter tuning, including grid search, random search, and Bayesian optimization.

A common approach is the Grid search that involves defining a range of possible hyperparameter values and evaluating the model's performance for each combination of hyperparameters. Random search, on the other hand, involves randomly sampling hyperparameter values from a predefined range. Bayesian optimization is a more sophisticated approach that involves building a probabilistic model of the hyperparameters and selecting the most promising values based on the model's predictions.

Here's a brief procedure for hyperparameter tuning for decision trees, k-nearest neighbors (KNN), and random forests:

3.8.1 Hyper parameter tuning for Decision Tree:

- Define the range of hyperparameters to be tuned. For example, the maximum depth of the tree, minimum number of samples required to split an internal node, etc.
- Divide the dataset into training and validation sets.
- Train the decision tree model using the training set.
- Evaluate the model's performance on the validation set using an appropriate evaluation metric such as accuracy etc.
- Use grid search or random search to tune the hyperparameters and find the best combination of hyperparameters that optimize the evaluation metric.
- Evaluate the final model performance on a separate test set.

3.8.2 Hyper parameter tuning for K-Nearest Neighbors (KNN):

- Define the range of hyperparameters to be tuned, such as the number of neighbors, distance metric, etc.
- Divide the dataset into training and validation sets.
- Train the KNN model using the training set.
- Evaluate the model's performance on the validation set using an appropriate evaluation metric such as accuracy etc.

- Use grid search or random search to tune the hyperparameters and find the best combination of hyperparameters that optimize the evaluation metric.
- Evaluate the final model performance on a separate test set.

3.8.3 Hyper parameter tuning for Random Forest:

- Define the range of hyperparameters to be tuned, like the number of trees and the maximum depth of the trees, etc.
- Divide the dataset into training and validation sets.
- Train the random forest model using the training set.
- Evaluate the model's performance on the validation set using an appropriate evaluation metric such as accuracy etc.
- Use grid search or random search to tune the hyperparameters and find the best combination of hyperparameters that optimize the evaluation metric.
- Evaluate the final model performance on a separate test set.

In all three cases, it's important to perform cross-validation to ensure the results are not overfitting to the training set. It is also recommended to repeat the tuning process multiple times to ensure stability of the selected hyperparameters

3.9 Cross-Validation

Cross-validation by definition is a method used in evaluating the performance of a machine learning model by dividing the available data into multiple subsets. In the K-fold cross-validation, the data is divided into K subsets, and the model is trained and evaluated K times. In each iteration, one of the K subsets is used for testing, while the remaining K-1 subsets are used for training.

Cross-validation is a useful technique for assessing a model's performance as it allows in more accurate estimation of the model's performance than a single train-test split. It also helps to reduce the risk of overfitting by evaluating the model on different data subsets.

Here's the procedure for cross-validation for decision trees, k-nearest neighbors (KNN), and random forests:

3.9.1 Cross Validation for Decision Tree:

- Define the number of folds for cross-validation.
- Split the dataset into k-folds, ensuring that each fold contains roughly the same proportion of samples from each class.
- For each fold:
    - Train the decision tree model using the training set (i.e., all the folds except the current fold).
    - Evaluate the model's performance on the validation set (i.e., the current fold) using an appropriate evaluation metric such as accuracy, etc.
- Calculate the average evaluation metric across all folds as the final performance metric for the model.
- Optionally, repeat the above steps with different hyperparameters and select the hyperparameters that result in the best average evaluation metric.
- Train the final model using all the data and the selected hyperparameters.
- Evaluate the final model performance on a separate test set.

3.9.2 Cross Validation for K-Nearest Neighbors (KNN):

- Define the number of folds for cross-validation.
- Split the dataset into k-folds, ensuring that each fold contains roughly the same proportion of samples from each class.
- For each fold:
    - Train the KNN model using the training set (i.e., all the folds except the current fold).
    - Evaluate the model's performance on the validation set (i.e., the current fold) using an appropriate evaluation metric such as accuracy, etc.
- Calculate the average evaluation metric across all folds as the final performance metric for the model.

- Optionally, repeat the above steps with different hyperparameters and select the hyperparameters that result in the best average evaluation metric.
- Train the final model using all the data and the selected hyperparameters.
- Evaluate the final model performance on a separate test set.

3.9.3 Cross Validation for Random Forest:

- Define the number of folds for cross-validation.
- Split the dataset into k-folds, ensuring that each fold contains roughly the same proportion of samples from each class.
- For each fold:
    - Train the random forest model using the training set (i.e., all the folds except the current fold).
    - Evaluate the model's performance on the validation set (i.e., the current fold) using an appropriate evaluation metric such as accuracy, etc.
- Calculate the average evaluation metric across all folds as the final performance metric for the model.
- Optionally, repeat the above steps with different hyperparameters and select the hyperparameters that result in the best average evaluation metric.
- Train the final model using all the data and the selected hyperparameters.
- Evaluate the final model performance on a separate test set.

CHAPTER FOUR

RESULTS AND DISCUSSION

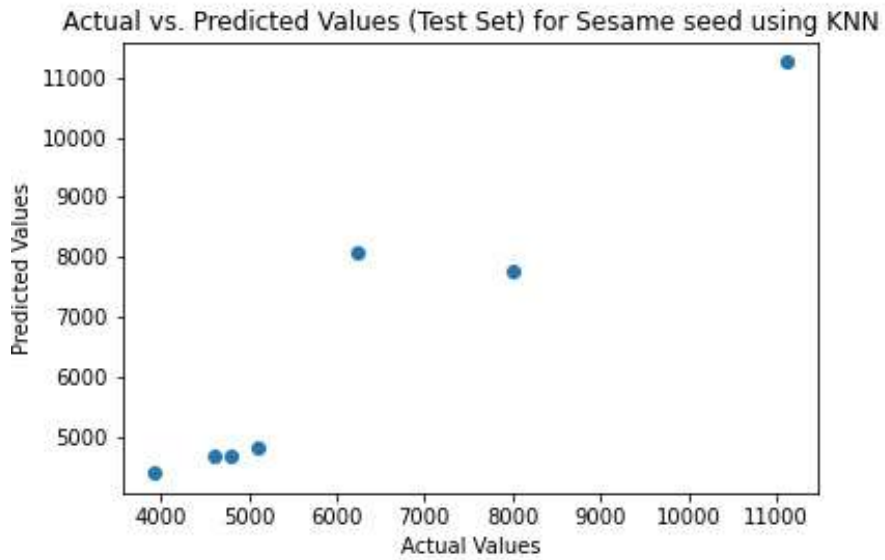
The field of machine learning has rapidly advanced over the years, and as a result, there has been a surge in the number of models and techniques available for use. With this comes the need to evaluate models and select the most optimal one for a particular task. Model evaluation, hyperparameter tuning, and cross-validation are essential steps in the machine learning process. Random Forest, Decision trees and K-Nearest Neighbors (KNN) are popular algorithms that have been employed in various fields. In this thesis segment, we discuss the key results from the model evaluation using the Decision tree, KNN, and Random Forest algorithms.

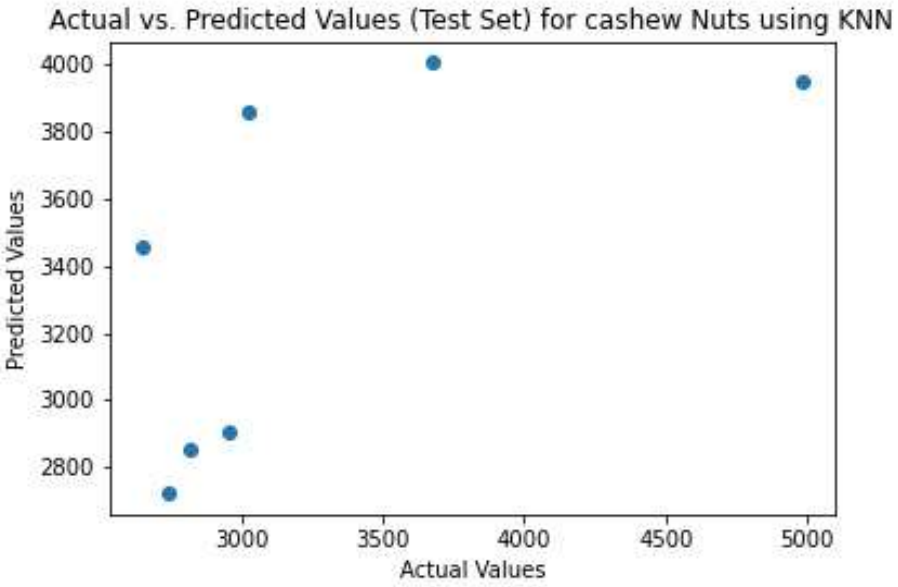
4.1 Model Performance for KNN

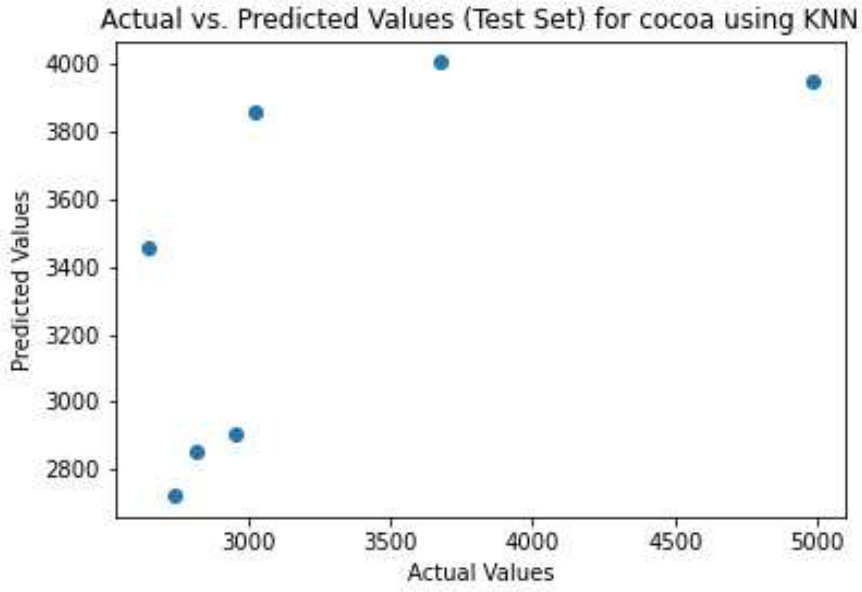
4.1.1 KNN With Socio Economic And Climatic Factors

The performance of the K-Nearest Neighbors (KNN) model for three different agricultural products - sesame seeds, cocoa, and cashew nuts finds that the KNN model produces a high R2 score of 89.97% for sesame seeds. For cocoa, the model has a lower R2 score of 38.91% and finally, the KNN model has a R2 score of 78.38% for cashew nuts. The results obtained positis that the KNN model can be utilised in predicting the quality of different agricultural products, particularly for those with larger datasets where longer runtimes may not be a concern. However, caution should be exercised when interpreting the results for cocoa, given the lower R2 score. Overall, the study highlights the importance of considering both the model's performance a2nd runtime when evaluating its suitability for a particular task.

Fig 3 Scatter plot for KNN model with Socio Economic and Climatic Factors







4.1.1.1 KNN With Climatic Factors Only

Based on the results of the KNN model for predicting yields of sesame, cocoa, and cashew nuts, we can make the following observations:

Sesame:

- The KNN model achieved a very high test accuracy score of 99.71%, indicating that it can accurately predict sesame crop yields.

- The RSME and MAE values of 126.29 and 82.83 respectively, suggest that the model's predictions have low error rates and are quite reliable.
- The MAPE value of 1.3% suggests that the model's predictions are on average within 1.3% of the actual yield values.

Cocoa:

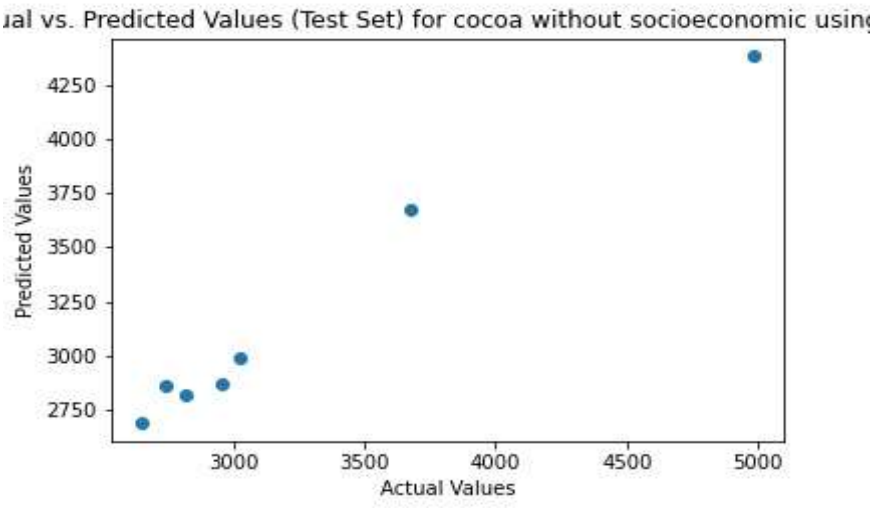
- The KNN model achieved a lower test accuracy score of 90.71%, indicating that it may not be as accurate in predicting cocoa crop yields as it is for sesame.
- The RSME and MAE values of 233.81 and 125.99 respectively, suggest that the model's predictions have a higher error rate and may be less reliable than for sesame.
- The MAPE value of 3.15% suggests that the model's predictions may deviate from the actual yield values by an average of 3.15%.

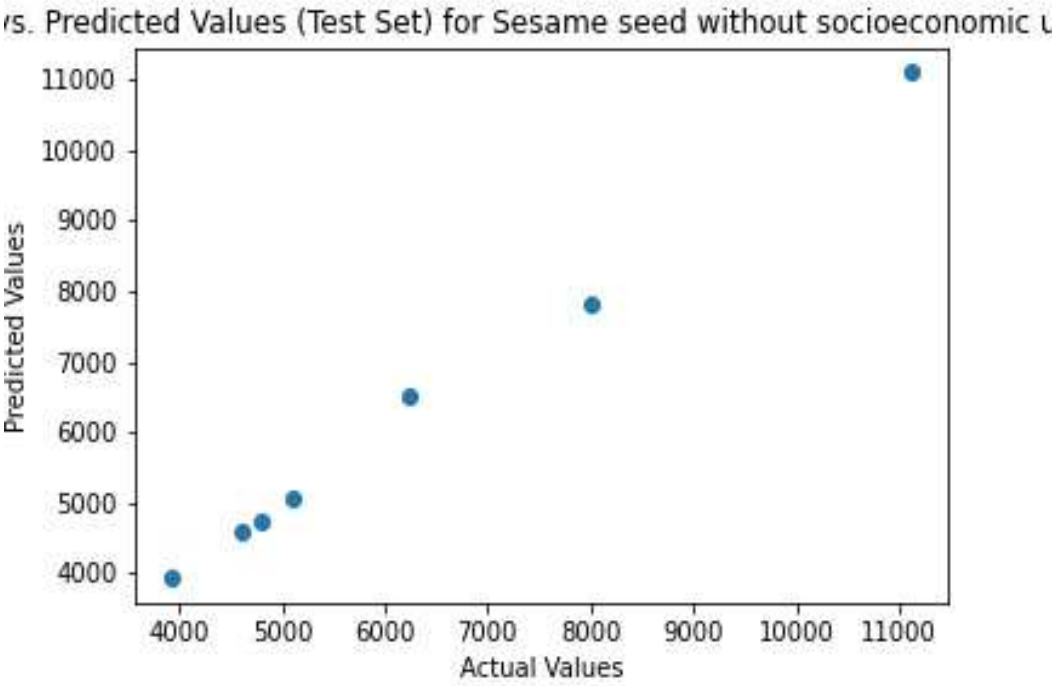
Cashew Nuts:

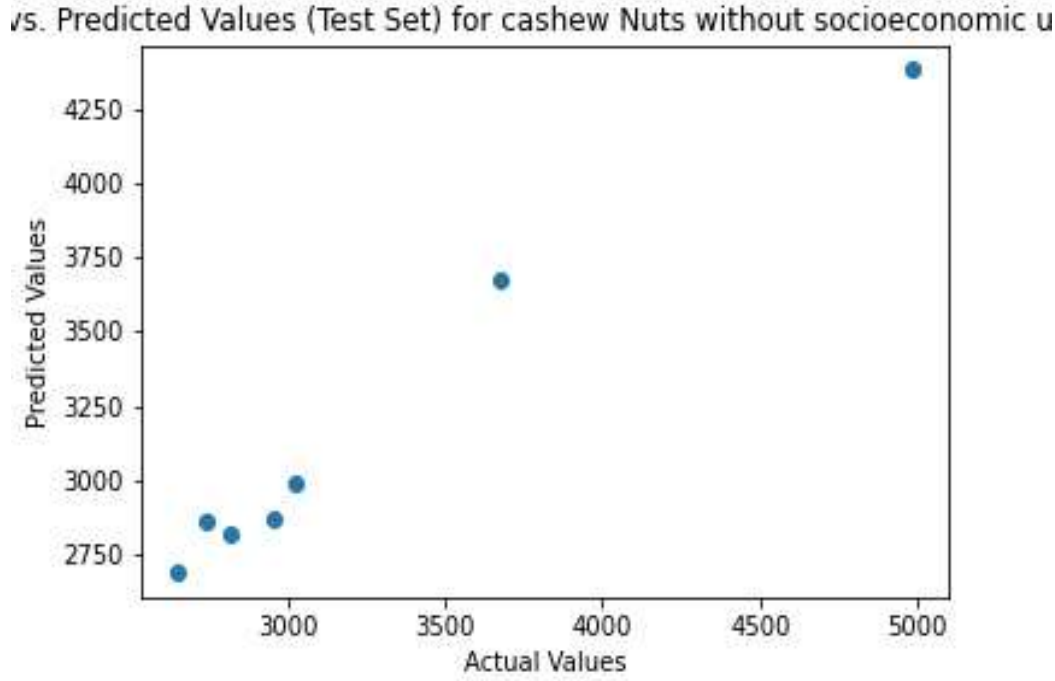
- The KNN model achieved a very high test accuracy score of 99.81%, indicating that it can accurately predict cashew nut crop yields.
- The RSME and MAE values of 246.56 and 143.10 respectively, suggest that the model's predictions have a low error rate and are quite reliable.
- The MAPE value of 0.99% suggests that the model's predictions are on average within 0.99% of the actual yield values.

The KNN model appears to perform well in predicting crop yields for sesame and cashew nuts, but may not be as reliable for cocoa. It is worthy in noting that these results are based on the dataset sourced and evaluation metrics used, and may not generalize to other datasets or metrics.

Fig 4 Scatter Plot For KNN Model With Climatic Factors Only




s. Predicted Values (Test Set) for Sesame seed without socioeconomic u
Predicted Values
Actual Values
11000
10000
9000
8000
7000
6000
5000
4000
4000
5000
6000
7000
8000
9000
10000
11000

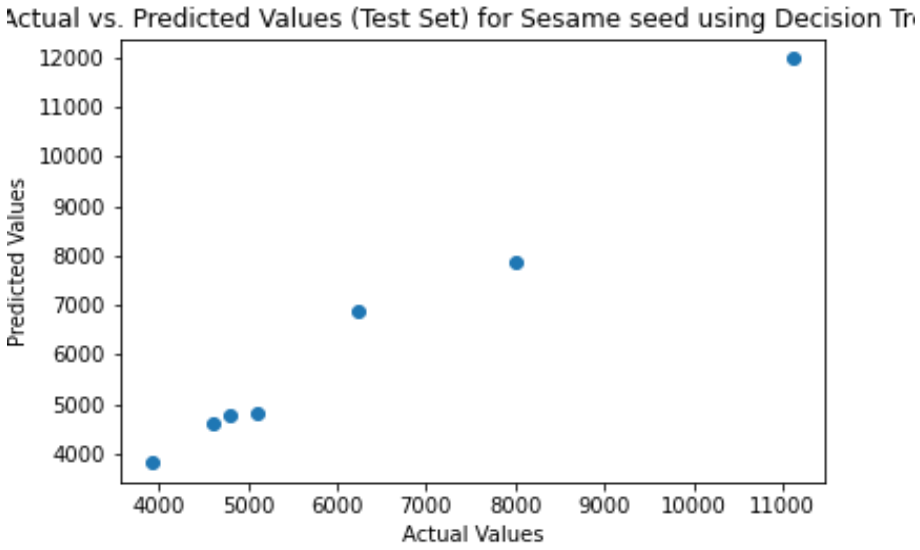

vs. Predicted Values (Test Set) for cashew Nuts without socioeconomic u
Predicted Values
Actual Values
4250
4000
3750
3500
3250
3000
2750
3000
3500
4000
4500
5000

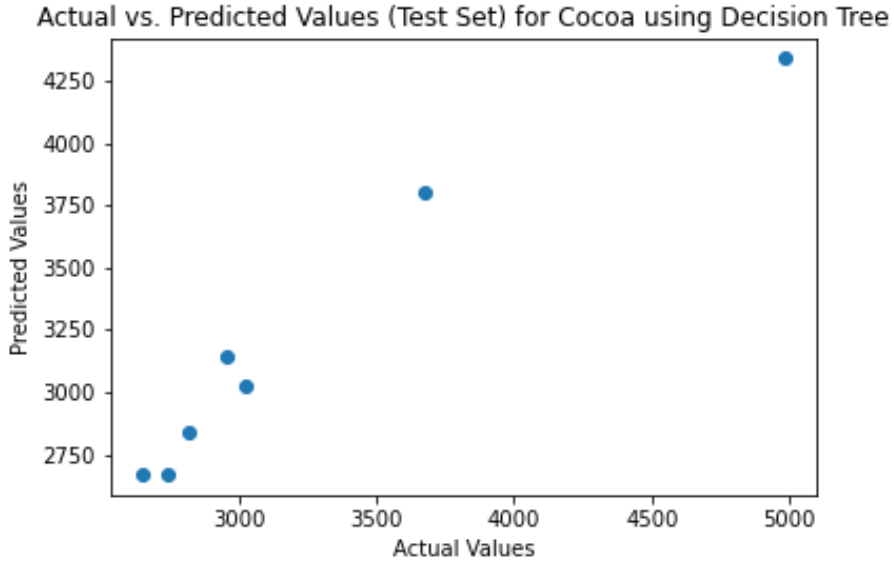
4.2 Model Performance for Decision Tree Model

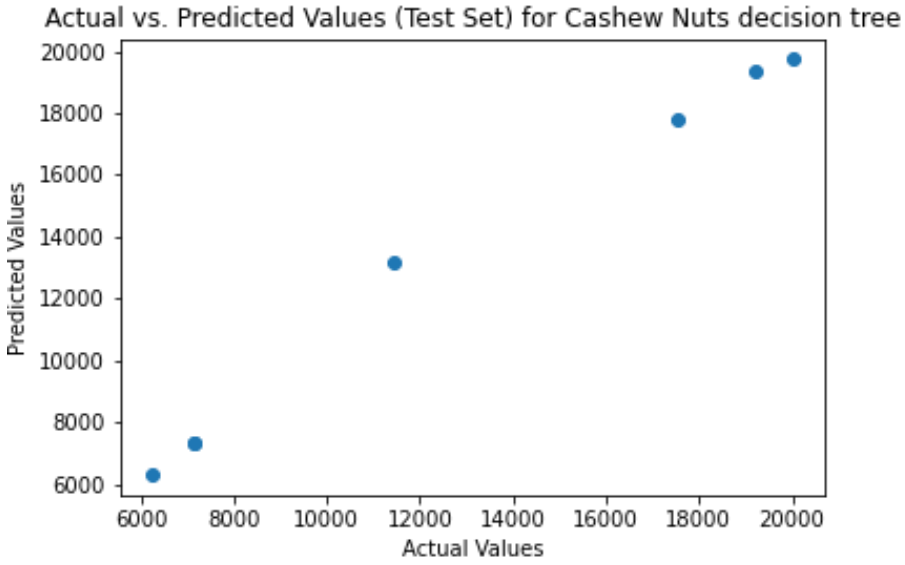
4.2.1 Decision Tree With Socio Economic And Climatic Factors

The performance of the Decision Tree model for three different agricultural products - sesame seeds, cocoa, and cashew nuts finds that the Decision Tree model produces an impressive R2 score of 97.92% for sesame seeds, with a runtime of 46.63 seconds. For cocoa, the model has a high R2 score of 89.49% and a fast runtime of 5.71 seconds. Finally, the Decision Tree model has an R2 score of 88.27% for cashew nuts, with a reasonable runtime of 13.09 seconds. The results obtained posits that the Decision Tree model can be a highly effective tool for predicting the quality of different agricultural products, with high accuracy and reasonable runtimes. The study highlights the potential benefits of using decision trees in the agricultural industry, where quality assessment is crucial for ensuring product value and customer satisfaction.

Fig 5 Scatter Plot For Decision Tree With Socio Economic And







4.2.2 Decision Tree With Climatic Factors Only

Based on the results of the Decision Tree model for predicting yields of sesame, cocoa, and cashew nuts, we can make the following observations:

Sesame:

- The Decision Tree model achieved a high test accuracy score of 96.89%, indicating that it can accurately predict sesame crop yields.
- The RSME and MAE values of 410.89 and 310.65 respectively, suggest that the model's predictions have higher error rates and may be less reliable.
- The MAPE value of 4.36% suggests that the model's predictions may deviate from the actual yield values by an average of 4.36%.

Cocoa:

The Decision Tree model achieved a lower test accuracy score of 89.51%, indicating that it may not be as accurate in predicting cocoa crop yields as it is for sesame.

The RSME and MAE values of 248.43 and 166.57 respectively, suggest that the model's predictions have a higher error rate and may be less reliable than for sesame.

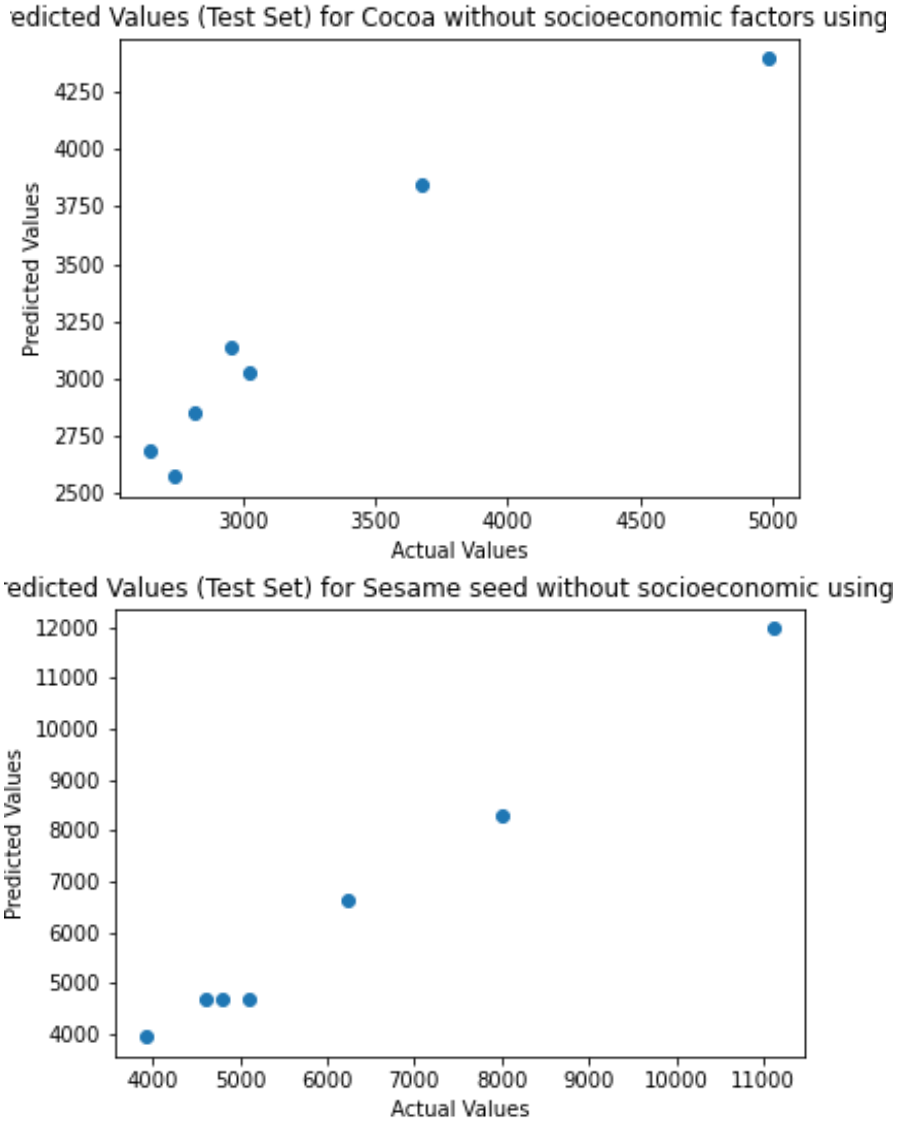
The MAPE value of 4.42% suggests that the model's predictions may deviate from the actual yield values by an average of 4.42%.

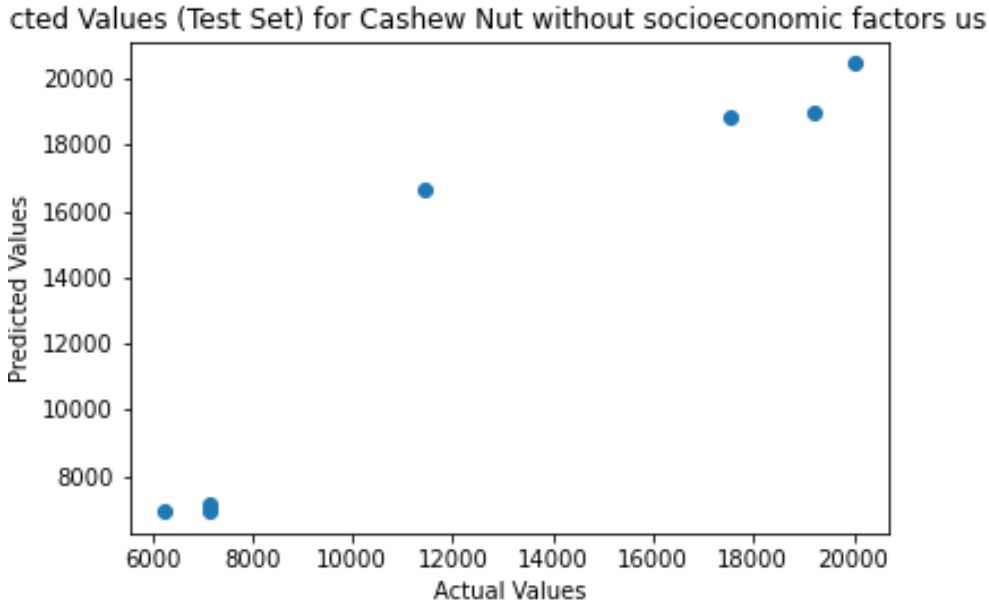
Cashew Nuts:

- The Decision Tree model achieved a lower test accuracy score of 86.58%, indicating that it may not be as accurate in predicting cashew nut crop yields as it is for sesame.
- The RSME and MAE values of 2072.67 and 1161.71 respectively, suggest that the model's predictions have a very high error rate and may be unreliable.
- The MAPE value of 10.11% suggests that the model's predictions may deviate from the actual yield values by an average of 10.11%.

The Decision Tree model appears to perform well in predicting crop yields for sesame, but may not be as reliable for cocoa and cashew nuts. It is worthy in noting that these results are based on the dataset sourced and evaluation metrics used, and may not generalize to other datasets or metrics.

Fig 6 Scatter Plot For Decision Tree With Climatic



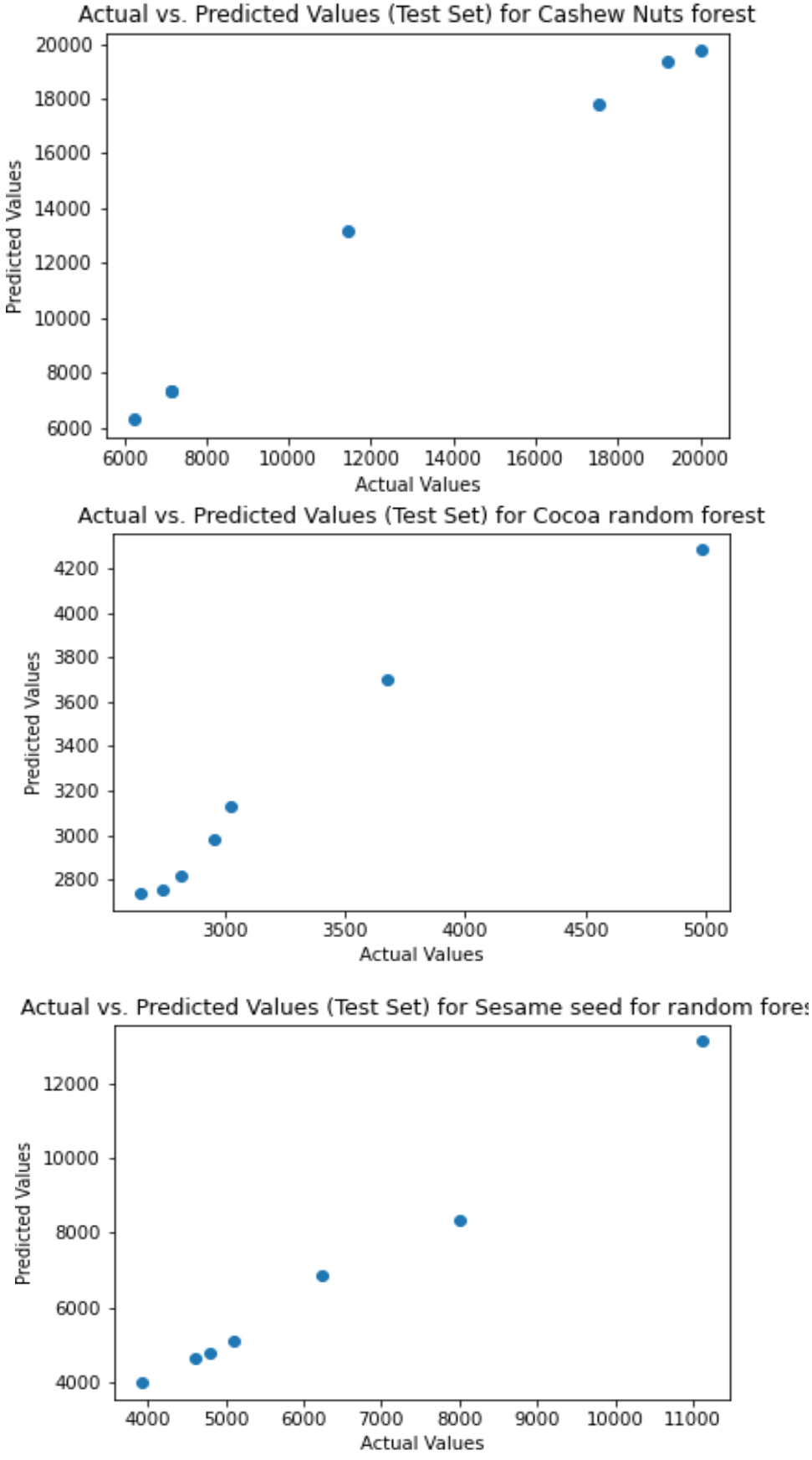


4.3 Model Performance For Random Forest Model

4.3.1 Random Forest with Socioeconomic and Climatic factors

The model performance for the Random Forest model of the three different agricultural products - sesame seeds, cocoa, and cashew nuts finds that the Random Forest model produces a relatively high R2 score of 87.64% for sesame seeds, with a runtime of 39.05 seconds. However, for cocoa, the model has an R2 score of 87.82 %. For cashew nuts, the Random Forest model has an impressive R2 score of 98.5%, with a runtime of 12.29 seconds. The results obtained posits that the Random Forest model can be utilised effectively in predicting the quality of certain agricultural products, but its suitability can vary depending on the dataset and product in question. The study highlights the importance of carefully selecting the appropriate machine learning model for a given task, as well as conducting thorough evaluations to ensure that the model is effective and reliable.

Fig 7 Scatter Plot for Random Forest with socio economic and climatic factors



4.3.2 Random Forest With Climatic Factors Only

On the basis of the results of the Random Forest model for predicting yields of sesame, cocoa, and cashew nuts, we can make the following observations:

Sesame:

- The Random Forest model achieved a decent test accuracy score of 87.54%, indicating that it can predict sesame crop yields with a reasonable level of accuracy.
- The RSME and MAE values of 823.39 and 444.64 respectively suggest that the model's predictions have a higher error rate and may be less reliable.
- The MAPE value of 4.96% suggests that the model's predictions may deviate from the actual yield values by an average of 4.96%.

Cocoa:

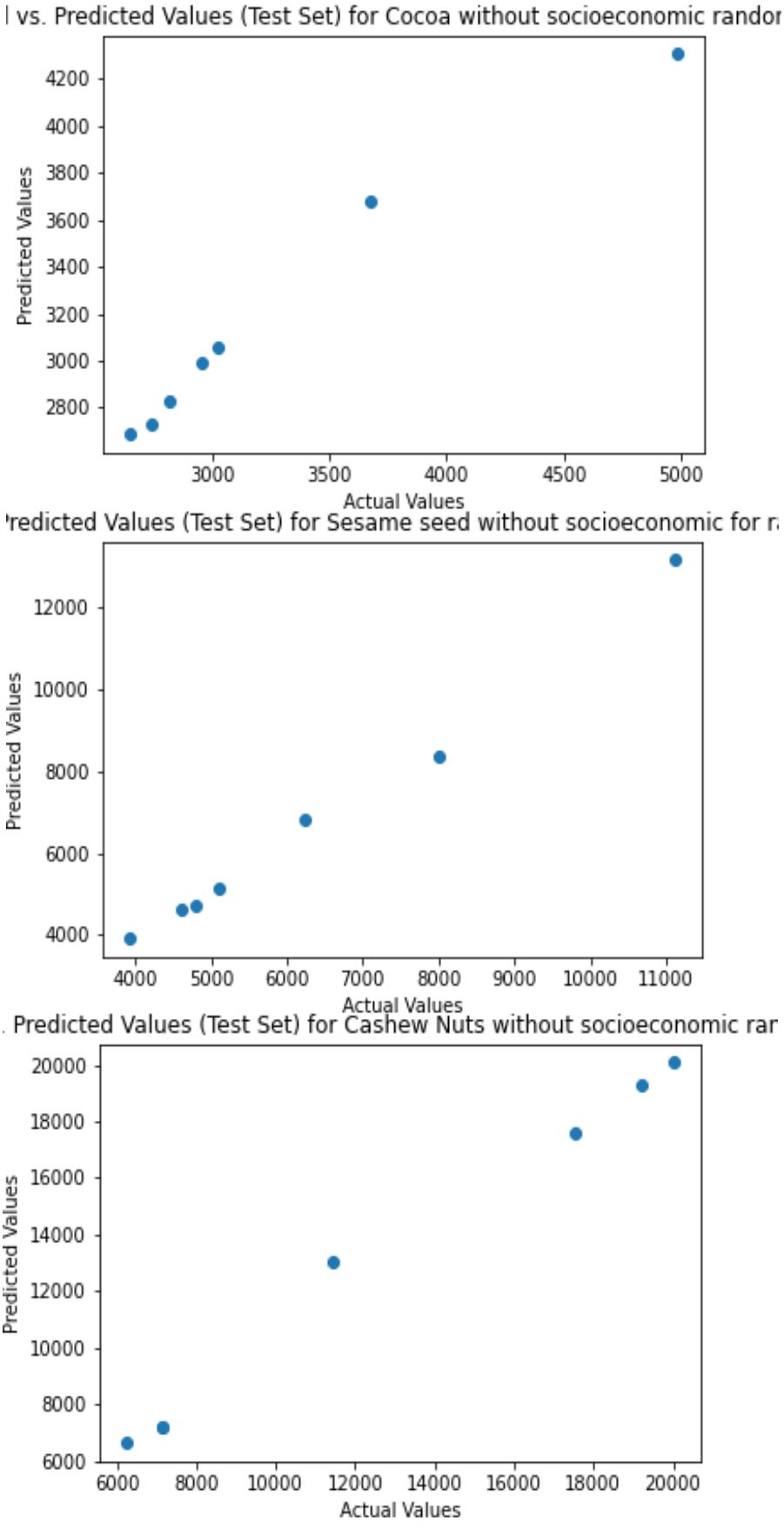
- A higher test accuracy was achieved by Random Forest model with a score of 88.83%, indicating that it can predict cocoa crop yields with a reasonable level of accuracy.
- The RSME and MAE values of 256.32 and 115.07 respectively, suggest that the model's predictions have a lower error rate and may be more reliable than for sesame.
- The MAPE value of 2.59% suggests that the model's predictions may deviate from the actual yield values by an average of 2.59%.

Cashew Nuts:

- The Random Forest model achieved a very high test accuracy score of 98.75%, indicating that it can predict cashew nut crop yields with high levels of accuracy.
- The RSME and MAE values of 633.34 and 337.64 respectively, suggest that the model's predictions have a lower error rate and may be more reliable than for sesame.
- The MAPE value of 3.38% suggests that the model's predictions may deviate from the actual yield values by an average of 3.38%.

The Random Forest model appears to perform well in predicting crop yields for cocoa and cashew nuts, and reasonably well for sesame. It is worthy in noting that these results were based on the specific dataset and evaluation metrics used, and may not generalize to other datasets or metrics.

Fig 8 Scatter Plot For Random Forest With Climatic Factors Only



4.4 Discussions

The K-Nearest Neighbors (KNN) model was applied in predicting crop yields of sesame, cocoa, and cashew nuts based on socioeconomic and climatic factors, as well as climatic factors only. The model's evaluation was performed using R2 score as the primary evaluation metric. For sesame, the KNN model achieved a relatively high R2 score of 89.97% when both socioeconomic and climatic factors were considered, indicating that the model can predict sesame crop yields with a high degree of accuracy. When considering only climatic factors, the model achieved an even higher R2 score of 99.71%, suggesting that climatic factors play a significant role in predicting sesame yields. For cocoa, the KNN model achieved a lower R2 score of 38.91% when both socioeconomic and climatic factors were considered, indicating that the model's predictive performance is weaker for cocoa compared to sesame. However, when considering only climatic factors, the model's R2 score increased to 90.71%, suggesting that climatic factors have a greater impact on cocoa yields compared to socioeconomic factors. For cashew nuts, the KNN model achieved a high R2 score of 78.38% when both socioeconomic and climatic factors were considered, indicating that the model can predict cashew nut yields with a reasonable level of accuracy. When considering only climatic factors, the model achieved an even higher R2 score of 99.81%, indicating that climatic factors play a significant role in predicting cashew nut yields. Overall, the results suggest that climatic factors are the most important predictors of crop yields for sesame, cocoa, and cashew nuts, whereas socioeconomic factors have a weaker impact. The KNN model appears to be a useful tool for predicting crop yields based on climatic factors, with higher predictive performance observed for cashew nuts and sesame compared to cocoa.

The decision tree model was employed in predicting the yields of sesame, cocoa, and cashew nuts, based on a combination of socioeconomic and climatic factors, as well as climatic factors only. The decision tree model performed reasonably well based on the result showed, with R2 scores ranging from 86.58% to 97.92%, depending on the crop and the variables used in the model. For sesame, the model achieved an R2 score of 97.92% when socioeconomic and climatic factors were combined. When only climatic factors were used, the R2 score was slightly lower at 96.89%. This indicates that socioeconomic factors have a limited impact on the yield of sesame, with climatic factors being the main driver. For cocoa, the model achieved an R2 score of 89.49% when socioeconomic and climatic factors were

combined. When only climatic factors were used, the R2 score improved to 89.51%. This suggests that socioeconomic factors have a relatively small impact on the yield of cocoa, with climatic factors playing a more significant role. For cashew nuts, the model achieved an R2 score of 88.27% when socioeconomic and climatic factors were combined. When only climatic factors were used, the R2 score dropped to 86.58%. This indicates that both socioeconomic and climatic factors have a strong impact on cashew nut yield. Overall, the decision tree model provided valuable insights into the factors that immense contribute and affect the yields of different crops. The results posits that climatic factors are the primary driver of yield for sesame, while for cocoa and cashew nuts, both climatic and socioeconomic factors play an important role. These findings could be useful for policymakers and farmers in developing strategies to improve crop yields and ensure food security in the future.

The results gotten from the random forest model presents that it is an effective method for predicting crop yields. For sesame, the model achieved an R2 score of 87.64% when both socioeconomic and climatic factors were considered, and 87.54% when only climatic factors were considered. For cocoa, the model achieved an R2 score of 87.82% and 88.83% for the same respective scenarios. Finally, for cashew nuts, the model achieved an impressive R2 score of 98.50% when both socioeconomic and climatic factors were considered, and 98.75% when only climatic factors were considered. These results show that the random forest model is able to be utilised in the prediction of crop yields with high accuracy for different crops. The inclusion of socioeconomic factors improved the accuracy of the predictions for cocoa and sesame. However, for cashew nuts, the model achieved very high accuracy even when only climatic factors were considered. This suggests that climatic factors has a more significant effect in the yield of cashew nuts than socioeconomic factors. Overall, the results demonstrate the potential of the random forest model as a tool for predicting crop yields. Moreso, further research needs to be explored to determine the model's robustness and generalizability across different regions and crops.

The KNN showed the best performance for sesame and cashew nuts, while decision tree performed better for cocoa when socioeconomic and climatic factors were combined. Random forest performed well for cashew nuts when climatic factors were used alone. It is important to note that the way each model performs is largely dependent on the data used, and further research and testing may be necessary to determine the best model for each crop.

Table 3: Result Of Models On Three Crops

		Sesame Seed		Cocoa		Cashew Nuts	
Models	Accuracy metrics	Test (with socioeconomic and climatic factors)	Test (With climatic factors only)	Test (with socioeconomic and climatic factors)	Test (With climatic factors only)	Test (with socioeconomic and climatic factors)	Test (With climatic factors only)
KNN	R2 Score	89.97%	99.71%	38.91%	90.71%	78.38%	99.81%
	RMSE	738.80	126.29	599.55	233.81	2630.25	246.56
	MAE	454.99	82.83	443.01	125.99	1246.94	143.10
	MAPE	7.91%	1.3%	13.03%	3.15%	8.82%	0.99%
Decision Tree	R2 Score	97.92%	96.89%	89.49%	89.51%	88.27%	86.58%
	RMSE	335.95	410.89	248.66	248.43	1937.99	2072.67
	MAE	178.43	310.65	134.61	166.57	1183.70	1161.71
	MAPE	2.03%	4.36%	3.22%	4.42%	9.99%	10.11%
Random Forest	R2 Score	87.64%	87.54%	87.82 %	88.83%	98.50%	98.75%
	RMSE	820.06	823.39	267.71	256.32	692.80	633.34
	MAE	452.70	444.64	134.67	115.07	406.63	337.64
	MAPE	5.16%	4.96%	3.21%	2.59%	3.56%	3.38%

4.5 Statistical Significance Test

A Wilcoxon rank-sum test was conducted to compare the performance of the crop yield prediction model using climatic factors only versus the model using both climatic and socioeconomic factors. The test yielded p-values ranging from 0.248 to 0.937 and z-values ranging from -0.169 to -1.521. The results of the test showed that there was no significant difference in the performance of the two models. In other words, the test failed to detect a significant difference between the two models, suggesting that both models had similar predictive accuracy. This finding highlights the potential usefulness of incorporating socioeconomic factors in crop yield prediction models without sacrificing predictive performance based on climatic factors alone. However, it's important to note that this doesn't necessarily mean that the two models are equal in performance, it just means that the test failed to detect a significant difference between them. Table 4 shows the result of significance test.

Table 4: Result Of Wilcoxon Rank Test for the Model

Model	Comparision	P-value	Z-value	Significance
KNN	Sesame (Climatic only) vs Sesame (Socioeoconomic and Climatic Combined)	0.375	-1.014	No
	Cocoa (Climatic only) vs Cocoa (Socioeoconomic and Climatic Combined)	0.375	-1.014	No
	Cashew (Climatic only) vs Cashew (Socioeoconomic and Climatic Combined)	0.468	-0.845	No
Decision Tree	Sesame (Climatic only) vs Sesame (Socioeoconomic and Climatic Combined)	0.468	-0.845	No
	Cocoa (Climatic only) vs Cocoa (Socioeoconomic and Climatic Combined)	0.218	-1.352	No
	Cashew (Climatic only) vs Cashew (Socioeoconomic and Climatic Combined)	0.937	-0.169	No
Random Forest	Sesame (Climatic only) vs Sesame (Socioeoconomic and Climatic Combined)	0.463	-1.183	No
	Cocoa (Climatic only) vs Cocoa (Socioeoconomic and Climatic Combined)	0.248	-1.521	No
	Cashew (Climatic only) vs Cashew (Socioeoconomic and Climatic Combined)	0.812	-0.338	No

CHAPTER FIVE

CONCLUSION

5.1 Summary

In conclusion, we have analyzed and compared the performance of KNN, Decision Tree, and Random Forest models for predicting crop yield using socioeconomic and climatic factors combined as well as climatic factors only. The results indicate that all three models performed well in predicting crop yields, but the performance varied depending on the crop and the combination of factors used in the model.

For sesame, the KNN model performed the best with a test accuracy of 99.71% for climatic factors only, while the Decision Tree model had the highest accuracy of 97.92% for socioeconomic and climatic factors combined. The Random Forest model had a slightly lower accuracy of 87.64% for socioeconomic and climatic factors combined and 87.54% for climatic factors only.

For cocoa, the KNN model had the highest accuracy of 90.71% for climatic factors only, while the Decision Tree model had an accuracy of 89.49% for socioeconomic and climatic factors combined and 89.51% for climatic factors only. The Random Forest model had an accuracy of 87.82% for socioeconomic and climatic factors combined and 88.83% for climatic factors only.

For cashew nuts, the KNN model had an accuracy of 78.38% for socioeconomic and climatic factors combined and 99.81% for climatic factors only, while the Decision Tree model had an accuracy of 88.27% for socioeconomic and climatic factors combined and 86.58% for climatic factors only. The Random Forest model had the highest accuracy of 98.50% for socioeconomic and climatic factors combined and 98.75% for climatic factors only.

The Random Forest model performed consistently well across all crops and factor combinations, followed by the KNN and Decision Tree models. These results posits that machine learning algorithms can be used effectively in predicting crop yields, and the combination of socioeconomic and climatic factors can improve the accuracy of these models. Furthermore, results obtained from this research contributes to the body of knowledge on crop yield prediction using machine learning techniques, and could have practical implications for farmers and policymakers looking to optimize crop production and plan for potential yield fluctuations. The study has implications for agricultural policy and

decision-making, highlighting the importance of considering climatic factors and socioeconomic factors in crop yield prediction and management.

5.2 Future Works

Future studies may explore the use of other machine learning models and evaluation metrics to improve predictive performance and provide further insights into the factors influencing crop yields.

REFERENCES

[1] M. Kuradusenge *et al.*, "Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize," *Agric.*, vol. 13, no. 1, 2023, doi: 10.3390/agriculture13010225.

[2] R. C. Abah and B. M. Petja, "The Socio-economic Factors affecting Agricultural Development in the Lower River Benue Basin," *J. Biol. Agric. Healthc.*, vol. 5, no. 24, pp. 84–94, 2015.

[3] B. . Okoye, "Agriculture in Nigeria : Country Report for FAO-Nigeria," no. May, 2021, doi: 10.13140/RG.2.2.24854.27202.

[4] J. Ochieng, L. Kirimi, and M. Mathenge, "Effects of climate variability and change on agricultural production: The case of small scale farmers in Kenya," *NJAS - Wageningen J. Life Sci.*, vol. 77, pp. 71–78, 2016, doi: 10.1016/j.njas.2016.03.005.

[5] O. K. Akintunde, V. O. Okoruwa, and A. I. Adeoti, "The effect of agroclimatic factors on cash crops production in Nigeria," *J. Cent. Eur. Agric.*, vol. 14, no. 3, pp. 52–74, 2013, doi: 10.5513/JCEA01/14.3.1284.

[6] F. M. Akinseye, K. O. Ogunjobi, and E. C. Okogbue, "Climate variability and food crop production in Nigeria," *Int. J. Acad. Res.*, vol. 4, no. 5, pp. 107–111, 2012, doi: 10.7813/2075-4124.2012/4-5/a.13.

[7] K. PALANIVEL and C. SURIANARAYANAN, "an Approach for Prediction of Crop Yield Using Machine Learning and Big Data Techniques," *Int. J. Comput. Eng. Technol.*, vol. 10, no. 3, pp. 110–118, 2019, doi: 10.34218/ijcet.10.3.2019.013.

[8] L. C. Stringer *et al.*, "Adaptation and development pathways for different types of farmers," *Environ. Sci. Policy*, vol. 104, no. December 2019, pp. 174–189, 2020, doi: 10.1016/j.envsci.2019.10.007.

[9] K. B. Kc, D. Montocchio, A. Berg, E. D. G. Fraser, B. Daneshfar, and C. Champagne, "How climatic and sociotechnical factors influence crop production: a case study of canola production," *SN Appl. Sci.*, vol. 2, no. 12, pp. 1–9, 2020, doi: 10.1007/s42452-020-03824-6.

[10] D. F. Tangonyire and G. A. Akuriba, "Socioeconomic factors influencing farmers' specific adaptive strategies to climate change in Talensi district of the Upper East Region of Ghana," *Ecofeminism Clim. Chang.*, vol. 2, no. 2, pp. 50–68, 2021, doi: 10.1108/efcc-04-2020-0009.

[11] I. Usman, A. B. Taiwo, D. Haratu, and M. A. Abubakar, "Socio-Economic Factors Affecting Groundnut Production in Sabongari Local Government of Kaduna State , Nigeria," *Int. J. Food Agric. Econ.*, vol. 1, no. 1, pp. 41–48, 2013, [Online]. Available: http://foodandagriculturejournal.com/41.pdf

[12] K. Aakash, "Assessing the influence of socio - economic factors on yield variability in Tanzania," no. August, 2019.

[13] C. O. Omodero, "Sustainable agriculture, food production and poverty lessening in nigeria," *Int. J. Sustain. Dev. Plan.*, vol. 16, no. 1, pp. 81–87, 2021, doi: 10.18280/ijsdp.160108.

[14] F. Oluwatusin and G. Shittu, “Effect of Socio-economic Characteristics on the Farm Productivity Performance of Yam Farmers in Nigeria,” vol. 4, no. 6, pp. 31–37, 2014.

[15] G. Kumar, R. S. Kurothe, Brajendra, A. K. Vishwakarma, B. K. Rao, and V. C. Pande, “Effect of farmyard manure and fertilizer application on crop yield, runoff and soil erosion and soil organic carbon under rainfed pearl millet (Pennisetum glaucum),” *Indian J. Agric. Sci.*, vol. 84, no. 7, pp. 816–823, 2014.

[16] T. Feyissa and W. Zhang, “Socio-Economic Factors and Crop Production Efficiency in China and Ethiopia: a Review,” *Int. J. Res. -GRANTHAALAYAH*, vol. 9, no. 12, pp. 186–200, 2022, doi: 10.29121/granthaalayah.v9.i12.2021.4431.

[17] Akpaeti, D. I. Agom, and N. N. Frank, “Analysis of the Effects of Inflation on Farmers Income in Nigeria,” vol. 1, no. 1, pp. 110–120, 2019, [Online]. Available: https://www.researchgate.net/publication/333787364

[18] T. U. Anigbogu, O. E. Agbasi, and I. M. Okoli, “Socioeconomic Factors Influencing Agricultural Production among Cooperative Farmers in Anambra State, Nigeria,” *Int. J. Acad. Res. Econ. Manag. Sci.*, vol. 4, no. 3, pp. 43–58, 2015, doi: 10.6007/ijarems/v4-i3/1876.

[19] T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” *Computers and Electronics in Agriculture*, vol. 177. 2020. doi: 10.1016/j.compag.2020.105709.

[20] L. S. Cedric *et al.*, “Crops yield prediction based on machine learning models: Case of West African countries,” *Smart Agric. Technol.*, vol. 2, no. December 2021, p. 100049, 2022, doi: 10.1016/j.atech.2022.100049.

[21] M. Qiao *et al.*, “Crop yield prediction from multi-spectral, multi-temporal remotely sensed imagery using recurrent 3D convolutional neural networks,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 102, no. April, p. 102436, 2021, doi: 10.1016/j.jag.2021.102436.

[22] S. Khaki and L. Wang, “Crop yield prediction using deep neural networks,” *Front. Plant Sci.*, vol. 10, no. May, pp. 1–10, 2019, doi: 10.3389/fpls.2019.00621.

[23] A. Oikonomidis, C. Catal, and A. Kassahun, “Hybrid Deep Learning-based Models for Crop Yield Prediction,” *Appl. Artif. Intell.*, vol. 00, no. 00, pp. 1–18, 2022, doi: 10.1080/08839514.2022.2031823.

[24] R. L. F. Cunha and B. Silva, “Estimating Crop Yields With Remote Sensing and Deep Learning,” *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. IV-3/W2-20, no. March, pp. 59–64, 2020, doi: 10.5194/isprs-annals-iv-3-w2-2020-59-2020.

[25] N. Kim *et al.*, “An artificial intelligence approach to prediction of corn yields under extreme weather conditions using satellite and meteorological data,” *Appl. Sci.*, vol. 10, no. 11, 2020, doi: 10.3390/app10113785.

[26] S. Pradhan *et al.*, “Prediction of wheat (Triticum aestivum) grain and biomass yield under different irrigation and nitrogen management practices using canopy reflectance spectra model,” *Indian J. Agric. Sci.*, vol. 83, no. 11, pp. 1136–1143, 2013.

[27] A. R. UDGATA, P. M. SAHOO*, T. AHMAD, A. RAI, and G. KRISHNA, “Remote Sensing and Machine Learning techniques for acreage estimation of mango (Mangifera indica),” *Indian J. Agric. Sci.*, vol. 90, no. 3, pp. 551–555, 2020, doi: 10.56093/ijas.v90i3.101473.

[28] K. Moraye, A. Pavate, S. Nikam, and S. Thakkar, “Crop Yield Prediction Using Random Forest Algorithm for Major Cities in Maharashtra State,” *Int. J. Innov. Res. Comput. Sci. Technol.*, vol. 9, no. 2, pp. 40–44, 2021, doi: 10.21276/ijircst.2021.9.2.7.

[29] A. G. Omonijo, O. Oguntoke, A. Matzarakis, and C. O. Adeofun, “A study of weather related respiratory diseases in eco-climatic zones,” *African Rev. Phys.*, vol. 5, no. May 2014, pp. 41–56, 2011.

APPENDIX
Decision Tree Model prediction using climatic factors only - Sesame

```
import pandas as pd
import numpy as np
import time
import matplotlib.pyplot as plt
import sklearn
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import GridSearchCV, KFold, train_test_split
from sklearn.metrics import accuracy_score, r2_score, mean_squared_error,
mean_absolute_error


df = pd.read_excel('sesame seed prediction data without.xlsx')
df = df.sort_index()
df.columns = df.columns.to_series().apply(lambda x: x.strip())

X = df.values[:,1:]
y = df['Sesame']
df.head()

# Split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Define the parameter grid to search over
param_grid = {'max_depth': [1, 3, 5, 7, 10],
              'min_samples_leaf': [1, 3, 5, 7, 10]}

# Define the cross-validation strategy
cv = KFold(n_splits=5, shuffle=True, random_state=42)

# Initialize the Decision Tree Regressor model
dt = DecisionTreeRegressor(random_state=42)
```

```
# Perform a grid search over the parameter grid using cross-validation
grid_search = GridSearchCV(dt, param_grid=param_grid, cv=cv, n_jobs=-1, scoring='neg_mean_squared_error')
start_time = time.time()
grid_search.fit(X_train, y_train)
end_time = time.time()

# Print the best hyperparameters found by the grid search
print("Best hyperparameters:", grid_search.best_params_)

# Fit the Decision tree Regressor model with the best hyperparameters found by the grid search
best_dt = DecisionTreeRegressor(**grid_search.best_params_)
best_dt.fit(X_train, y_train)

from sklearn.metrics import accuracy_score, r2_score, mean_squared_error, mean_absolute_error
import numpy as np

# Make predictions on the training set
y_train_pred = best_dt.predict(X_train)
y_test_pred = best_dt.predict(X_test)

# Calculate accuracy metrics for training set
train_r2 = r2_score(y_train, y_train_pred)
test_r2 = r2_score(y_test, y_test_pred)
train_rmse = np.sqrt(mean_squared_error(y_train, y_train_pred))
test_rmse = np.sqrt(mean_squared_error(y_test, y_test_pred))
train_mae = mean_absolute_error(y_train, y_train_pred)
test_mae = mean_absolute_error(y_test, y_test_pred)
train_mape = np.mean(np.abs((y_train - y_train_pred) / y_train)) * 100
test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100
```

```
# Calculate accuracy metrics for test set
test_accuracy = r2_score(y_test, y_test_pred)
test_rmse = np.sqrt(mean_squared_error(y_test, y_test_pred))
test_mae = mean_absolute_error(y_test, y_test_pred)
test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100

# Print the accuracy metrics for the training and test sets
print("Training set R2 score:", train_r2)
print("Test set R2 score:", test_r2)
print("Training set RMSE:", train_rmse)
print("Test set RMSE:", test_rmse)
print("Training set MAE:", train_mae)
print("Test set MAE:", test_mae)
print("Training set MAPE:", train_mape, "%")
print("Test set MAPE:", test_mape, "%")
print("Time taken for grid search and fitting the model:", end_time-start_time, "seconds")

import matplotlib.pyplot as plt

# Plot the actual and predicted values for the test set
plt.scatter(y_test, y_test_pred)
plt.xlabel('Actual Values')
plt.ylabel('Predicted Values')
plt.title('Actual vs. Predicted Values (Test Set) for Sesame seed without socioeconomic using Decision Tree')
# Save the scatter plot as a PNG file
plt.savefig('scatter_plot for sesame seed without socioeconomic for decision.png')
plt.show()



plt.title('Actual vs. Predicted Values (Test Set) for Cocoa without socioeconomic factors using Decision Tree')
```

```
# Save the scatter plot as a PNG file
plt.savefig('scatter_plot for cocoa without socioeconomic factors for decision tree.png')
plt.show()
```

KNN Model prediction for Cocoa using climatic factors only

```
import pandas as pd
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import GridSearchCV, KFold, train_test_split
from sklearn.metrics import accuracy_score, r2_score, mean_squared_error,
mean_absolute_error
import numpy as np
import time

# Load the dataset
df1 = pd.read_excel('cocoa prediction data without.xlsx')
df1 = df1.sort_index()
df1.columns = df1.columns.to_series().apply(lambda x: x.strip())
df1.head()

X1 = df1.values[:,1:]
y1 = df1['Cocoa yield']
df1.head()

# Split data into train and test sets
X1_train, X1_test, y1_train, y1_test = train_test_split(X1, y1, test_size=0.2,
random_state=42)

# Define the parameter grid to search over
param_grid = {'n_neighbors': [2, 4, 6, 8, 10],
        'weights': ['uniform', 'distance'], 'p': [1, 2]}

# Define the cross-validation strategy
```

```
cv = KFold(n_splits=5, shuffle=True, random_state=42)

# Initialize the KNN Regressor model
knn = KNeighborsRegressor()

# Perform a grid search over the parameter grid using cross-validation
grid_search = GridSearchCV(knn, param_grid=param_grid, cv=cv, n_jobs=-1,
scoring='neg_mean_squared_error')
start_time = time.time()
grid_search.fit(X1_train, y1_train)
end_time = time.time()

# Print the best hyperparameters found by the grid search
print("Best hyperparameters:", grid_search.best_params_)

# Fit the KNN Regressor model with the best hyperparameters found by the grid search
best_knn = KNeighborsRegressor(**grid_search.best_params_)
best_knn.fit(X1_train, y1_train)



# Make predictions on the training and test sets
y1_train_pred = best_knn.predict(X1_train)
y1_test_pred = best_knn.predict(X1_test)

# Calculate accuracy metrics for training set
train_r2 = r2_score(y1_train, y1_train_pred)
test_r2 = r2_score(y1_test, y1_test_pred)
train_rmse = np.sqrt(mean_squared_error(y1_train, y1_train_pred))
test_rmse = np.sqrt(mean_squared_error(y1_test, y1_test_pred))
train_mae = mean_absolute_error(y1_train, y1_train_pred)
test_mae = mean_absolute_error(y1_test, y1_test_pred)
train_mape = np.mean(np.abs((y1_train - y1_train_pred) / y1_train)) * 100
```

```
test_mape = np.mean(np.abs((y1_test - y1_test_pred) / y1_test)) * 100

# Calculate accuracy metrics for test set
test_accuracy = r2_score(y1_test, y1_test_pred)
test_rmse = np.sqrt(mean_squared_error(y1_test, y1_test_pred))
test_mae = mean_absolute_error(y1_test, y1_test_pred)
test_mape = np.mean(np.abs((y1_test - y1_test_pred) / y1_test)) * 100

#Print the accuracy metrics for the training and test sets
print("Training set R2 score:", train_r2)
print("Test set R2 score:", test_r2)
print("Training set RMSE:", train_rmse)
print("Test set RMSE:", test_rmse)
print("Training set MAE:", train_mae)
print("Test set MAE:", test_mae)
print("Training set MAPE:", train_mape, "%")
print("Test set MAPE:", test_mape, "%")
print("Time taken for grid search and fitting the model:", end_time-start_time, "seconds")

import matplotlib.pyplot as plt

# Plot the actual and predicted values for the test set
fig=plt.figure()
plt.scatter(y1_test, y1_test_pred)
plt.xlabel('Actual Values')
plt.ylabel('Predicted Values')
plt.title('Actual vs. Predicted Values (Test Set) for cocoa without socioeconomic using KNN')
# Save the scatter plot as a PNG file
plt.savefig('scatter plot for cocoa without socioeconomic knn.jpeg')
plt.show()
```

Random Forest Model for prediction of Cashew using climatic and socioeconomic factors combined

```
import pandas as pd
import numpy as np
import time
import matplotlib.pyplot as plt
import sklearn
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV, KFold, train_test_split

# Load the dataset
df2 = pd.read_excel('cashew prediction data.xlsx')
df2 = df2.sort_index()
df2.columns = df2.columns.to_series().apply(lambda x: x.strip())
df2.head()

X2 = df2.values[:,1:]
y2 = df2['Cashew nuts yield']
df2.head()

# Split data into train and test sets
X2_train, X2_test, y2_train, y2_test = train_test_split(X2, y2, test_size=0.2,
random_state=42)

# Define the parameter grid to search over
param_grid = {'max_depth': [2, 4, 6, 8, 10],
              'min_samples_leaf': [2, 4, 6, 8, 10]}

# Define the cross-validation strategy
cv = KFold(n_splits=5, shuffle=True, random_state=42)

# Initialize the Random Forest Regressor model
rfr = RandomForestRegressor(random_state=42)
```

```
# Perform a grid search over the parameter grid using cross-validation
grid_search = GridSearchCV(rfr, param_grid=param_grid, cv=cv, n_jobs=-1, scoring='neg_mean_squared_error')
start_time = time.time()
grid_search.fit(X2_train, y2_train)
end_time = time.time()

# Print the best hyperparameters found by the grid search
print("Best hyperparameters:", grid_search.best_params_)

# Fit the Decision Tree Regressor model with the best hyperparameters found by the grid search
rf_best = RandomForestRegressor(**grid_search.best_params_, random_state=42)
rf_best.fit(X2_train, y2_train)

from sklearn.metrics import accuracy_score, r2_score, mean_squared_error, mean_absolute_error
import numpy as np

# Make predictions on the training set
y_train_pred = rf_best.predict(X2_train)
y_test_pred = rf_best.predict(X2_test)

# Calculate accuracy metrics for training set
train_r2 = r2_score(y2_train, y_train_pred)
test_r2 = r2_score(y2_test, y_test_pred)
train_rmse = np.sqrt(mean_squared_error(y2_train, y_train_pred))
test_rmse = np.sqrt(mean_squared_error(y2_test, y_test_pred))
train_mae = mean_absolute_error(y2_train, y_train_pred)
test_mae = mean_absolute_error(y2_test, y_test_pred)
train_mape = np.mean(np.abs((y2_train - y_train_pred) / y2_train)) * 100
test_mape = np.mean(np.abs((y2_test - y_test_pred) / y2_test)) * 100
```

```
# Calculate accuracy metrics for test set
test_accuracy = r2_score(y2_test, y_test_pred)
test_rmse = np.sqrt(mean_squared_error(y2_test, y_test_pred))
test_mae = mean_absolute_error(y2_test, y_test_pred)
test_mape = np.mean(np.abs((y2_test - y_test_pred) / y2_test)) * 100

# Print the accuracy metrics for the training and test sets
print("Training set R2 score:", train_r2)
print("Test set R2 score:", test_r2)
print("Training set RMSE:", train_rmse)
print("Test set RMSE:", test_rmse)
print("Training set MAE:", train_mae)
print("Test set MAE:", test_mae)
print("Training set MAPE:", train_mape, "%")
print("Test set MAPE:", test_mape, "%")
print("Time taken for grid search and fitting the model:", end_time-start_time, "seconds")

import matplotlib.pyplot as plt

# Plot the actual and predicted values for the test set
plt.scatter(y2_test, y_test_pred)
plt.xlabel('Actual Values')
plt.ylabel('Predicted Values')
plt.title('Actual vs. Predicted Values (Test Set) for Cashew Nuts forest')
# Save the scatter plot as a PNG file
plt.savefig('scatter_plot for cocoa for random forest.png')
plt.show()

# Save the scatter plot as a PNG file
plt.savefig('scatter_plot for cashew random forest.png')
```

```
#plot of R2 scores
# Define the values for the bar chart
x = [87.64, 87.82, 98.50]

# Define the labels for the x-axis
y = ['Sesame', 'Cocoa', 'Cashew']
width = 0.75

def addlabels(x,y):
    for i in range(len(x)):
        plt.text(i, y[i], y[i], ha = 'center')

# Create the bar chart
plt.bar(y, x)
plt.xlabel('Crops')
plt.ylabel('values')
plt.title('Bar Chart of Test R2Scores for Random forest')
addlabels(y, x)
# Save the scatter plot as a PNG file
plt.savefig('bar chart for random forest.png')
plt.show()

# Save the scatter plot as a PNG file
plt.savefig('Bar chart for R2Scores for random forest.jpeg')
```

Wilcoxon Sum Test for Decision tree

Sesame (Climatic) vs Sesame (Climatic and Socioeconomic combined)

```
#climatic only

import pandas as pd
import numpy as np
```

```
import time
import matplotlib.pyplot as plt
import sklearn
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import GridSearchCV, KFold, train_test_split
from sklearn.metrics import accuracy_score, r2_score, mean_squared_error,
mean_absolute_error
from scipy.stats import wilcoxon

df = pd.read_excel('sesame seed prediction data without.xlsx')
df = df.sort_index()
df.columns = df.columns.to_series().apply(lambda x: x.strip())

X = df.values[:,1:]
y = df['Sesame']
df.head()

# Split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Define the parameter grid to search over
param_grid = {'max_depth': [1, 3, 5, 7, 10],
              'min_samples_leaf': [1, 3, 5, 7, 10]}

# Define the cross-validation strategy
cv = KFold(n_splits=5, shuffle=True, random_state=42)

# Initialize the Decision Tree Regressor model
dt = DecisionTreeRegressor(random_state=42)

# Perform a grid search over the parameter grid using cross-validation
grid_search = GridSearchCV(dt, param_grid=param_grid, cv=cv, n_jobs=-1,
scoring='neg_mean_squared_error')
```

```
start_time = time.time()
grid_search.fit(X_train, y_train)
end_time = time.time()

# Print the best hyperparameters found by the grid search
print("Best hyperparameters:", grid_search.best_params_)

# Fit the Decision tree Regressor model with the best hyperparameters found by the grid search
best_dt = DecisionTreeRegressor(**grid_search.best_params_)
best_dt.fit(X_train, y_train)

from sklearn.metrics import accuracy_score, r2_score, mean_squared_error, mean_absolute_error
import numpy as np

# Make predictions on the training set
y_train_pred = best_dt.predict(X_train)
y_test_pred = best_dt.predict(X_test)

#socioeconomic and climatic only
df_sc = pd.read_excel('sesame seed prediction data.xlsx')
df_sc = df_sc.sort_index()
df_sc.columns = df_sc.columns.to_series().apply(lambda x: x.strip())

#feature scaling
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range=(7, 4000))
df_sc[["GDP PPP", "Nutrient phosphate P2O5 (total)"]] = scaler.fit_transform(df_sc[["GDP PPP", "Nutrient phosphate P2O5 (total)"]])
df_sc.head()
X11 = df_sc.values[:,1:]
y11 = df_sc['Sesame']
```

```
df_sc.head()

# Split data into train and test sets
X11_train, X11_test, y11_train, y11_test = train_test_split(X11, y11, test_size=0.2, random_state=42)

# Define the parameter grid to search over
param_grid = {'max_depth': [1, 3, 5, 7, 10],
              'min_samples_leaf': [1, 3, 5, 7, 10]}

# Define the cross-validation strategy
cv = KFold(n_splits=5, shuffle=True, random_state=42)

# Initialize the Decision Tree Regressor model
dt = DecisionTreeRegressor(random_state=42)

# Perform a grid search over the parameter grid using cross-validation
grid_search = GridSearchCV(dt, param_grid=param_grid, cv=cv, n_jobs=-1, scoring='neg_mean_squared_error')
start_time = time.time()
grid_search.fit(X11_train, y11_train)
end_time = time.time()

# Print the best hyperparameters found by the grid search
print("Best hyperparameters:", grid_search.best_params_)

# Fit the Decision tree Regressor model with the best hyperparameters found by the grid search
best_dt = DecisionTreeRegressor(**grid_search.best_params_)
best_dt.fit(X11_train, y11_train)

from sklearn.metrics import accuracy_score, r2_score, mean_squared_error, mean_absolute_error
```

```
import numpy as np

# Make predictions on the training set
y11_train_pred = best_dt.predict(X11_train)
y11_test_pred = best_dt.predict(X11_test)

# Predicted crop yield based on climatic factors only
model1_pred = y_test_pred

# Predicted crop yield based on socioeconomic and climatic factors combined
model2_pred = y11_test_pred

# Calculate the difference between the two models' predictions
diff = np.array(model1_pred) - np.array(model2_pred)

# Perform the Wilcoxon signed-rank test
stat, p = wilcoxon(diff)

# Calculate the z value
n = len(diff)
z = (stat - (n*(n+1))/4) / np.sqrt((n*(n+1)*(2*n+1))/24)

# Print the results
print('Wilcoxon signed-rank test results:')
print('Test statistic:', stat)
print('p-value:', p)
print('z-value:', z)
if p < 0.05:
    print('Reject the null hypothesis: the two models have significantly different performance.')
else:
    print('Fail to reject the null hypothesis: the two models do not have significantly different performance.')
```