

SMART MANGO – A SMART SYSTEM FOR MANGO PLANTATION MANAGEMENT

Project ID: 23-309

Final Thesis (Individual)

Niroshani A. – IT20103354

BSc (Hons) Degree in Information Technology specializing in Information
Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

February 2023

Smart Mango – A Smart System for Mango Plantation
Management
Yield Prediction

Project ID: 23-309

Final Thesis (Individual)

Niroshani A. – IT20103354

Supervisor: Ms. Hansika Mahaadikara

Co-Supervisor: Ms. Shashika Lokuliyana

Dissertation Submitted in Partial Fulfillment of the Requirements for the BSc (Hons) in
Information Technology

Department of Information Technology
Sri Lanka Institute of Information Technology
Sri Lanka

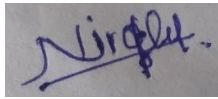
February 2023

Declaration

I declare that this is my own work, and this Thesis does not incorporate without acknowledgment any material previously submitted for a degree or Diploma in any other University or institute of higher learning, and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology, the non-exclusive right to reproduce and distribute my Thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:



Date: 09/09/2023

The above candidate has carried out this research thesis for the Degree of Bachelor of Science (honors) in Information Technology (Specializing in Information Technology) under my supervision.

Signature of the supervisor:

Date:

Acknowledgment

I would like to express my special thanks and gratitude to those who helped me continue my 4th-year research project. First, I sincerely thank our Research Project Supervisor Ms. Hansika Mahaadikara, who guided and encouraged me to continue this project successfully. I am also grateful for the guidance given by the AIMS panel through their advice for our topic assessment, I would like to thank our Co-Supervisor Ms. Shashika Lokuliyana for her valuable advice in fulfilling our work more efficiently.

I also acknowledge my Research Project leader and other group members who helped me. Finally, I would like to give my heartfelt gratitude to my parents and friends who helped me to make this project successful.

Table of Content

DECLARATION	I
ACKNOWLEDGMENT	II
LIST OF FIGURES	IV
LIST OF TABLE	V
LIST OF ABBREVIATION	VI
ABSTRACTION	1
1. INTRODUCTION	2
1.1 <i>Background</i>	3
1.2 <i>Literature Review</i>	7
1.3 <i>Research Gap</i>	9
2. RESEARCH PROBLEM	11
3. RESEARCH OBJECTIVE	14
3.1 <i>Main Objective</i>	14
3.2 <i>Sub-objective 1:</i>	14
3.3 <i>Sub-objective 2:</i>	14
3.4 <i>Sub-objective 3:</i>	14
3.5 <i>Sub-objective 4:</i>	14
4. METHODOLOGY	15
4.1 <i>Testing and Implementation</i>	21
5. RESULTS AND DISCUSSION	22
5.1 <i>Results</i>	22
5.2 <i>Discussion</i>	24
6. CONCLUSION	25
REFERENCES.....	26
APPENDICES.....	31

List of Figures

Figure 1: Application Overview	3
Figure 2: Overall Architecture	18
Figure 3: Data set	19
Figure 4: Synthetic data row	20
Figure 5: Learning Curve.....	22
Figure 6: Training Accuracy	23
Figure 7: Mobile app Login	31
Figure 8: Mobile app dashboard	32
Figure 9: Mobile app yield prediction.....	33
Figure 10: Mobile app Log out	34

List of Table

Table 1: Research Gap 10

Table 2: Testing results 21

List of Abbreviation

Abbreviation	Description
pH	Potential of hydrogen
GBM	Gradient Boosting Machine
SVM	Support Vector Machine
ML	Machine Learning
MSE	Mean Squared Error

Abstraction

Sri Lanka has a large mango farming industry. The average yield of mangoes per acre in Sri Lanka, according to the latest statistics available, is around 7-8 metric tons per year. The overall mango production in Sri Lanka for the 2020–2021 season was estimated to be around 39,862 metric tons, down from 49,535 metric tons the previous season previously [1]. The yield and profitability of the farmer depend greatly on the accuracy of the harvest forecast. Most farmers prefer to refer to past data on the yield and take steps to increase the yield. Farmers nevertheless continue to experience issues with yield and income due to the forecast's unreliability. One of the industry's biggest problems is the lack of a reliable method for accurate prediction for a variety of mangoes. As a solution for this issue, a forecasting model could be presented to Sri Lankan mango farmers to assist them in addressing the decline in output and predicting mango yield for future seasons. That utilizes different factors including Soil pH, Soil Moisture, temperature, humidity, Rainfall, Light Exposure, Life Span, pesticides, and diseases to predict the harvest. Mango producers could make better choices about their methods of mango cultivation, such as how much fertilizer, water, and pesticides to use on their trees, by utilizing this strategy. Additionally, they could forecast the best time to harvest their mangoes using the prediction model, which can influence the sweetness and flavor of the produce. Overall, Smart Mango can assist Sri Lankan mango producers in making the most of their resources, increasing yields and profitability despite changes in the weather and other environmental variables.

KEYWORDS: Smart Mango, Machine Learning, Prediction Model, Temperature, Humidity, Soil Moisture, Pesticides, Diseases.

1. INTRODUCTION

The production of mangoes in Sri Lanka is a very successful industry that provides a significant contribution to the agricultural environment of the nation. Despite this, the sector must contend with a recurrent challenge, namely the trustworthiness of mango production estimates. The feasibility and profitability of mango cultivation are intricately linked to the accuracy of harvest projections, which in turn influences crucial decisions that are made by farmers. Mangoes are only cultivated in regions of the globe with tropical weather. Even though historical data has been the go-to resource for many producers trying to boost their yields, the chronic unpredictability of mango harvests continues to be a challenge for the industry. This is a problem even though previous data has been available.

The harvesting of about 39,862 metric tons of mangoes in Sri Lanka during the season of 2020-2021 was followed by a decrease in output as compared to the previous year. When compared to the sum of 49,535 metric tons produced during the prior season, this figure represents a decrease in output. The unpredictability of mango production, which is often subject to swings that are impacted by a different set of climatic factors, highlights the need to discover a system that is more trustworthy in terms of forecasting. Mango production is frequently susceptible to swings that are influenced by a distinctive collection of climatic variables.

This thesis tries to provide a solution to this big problem by putting forth a forecasting model that was constructed specifically for Sri Lankan mango farmers and giving it the catchy label "Smart Mango." This model was created to help the farmers better predict future mango production. The number of mangoes that will be grown in the world in the years to come may be estimated with the help of this model.

The Smart Mango model makes use of cutting-edge techniques of machine learning, with a substantial focus on linear regression as the primary way of analysis. These cutting-edge methods of machine learning are included in the model. This is done so that an accurate prognosis may be made about the quantity of mangoes that will be produced throughout the next seasons. It is important to take into consideration several factors, including the pH of the soil, the amount of moisture that is contained within the soil, the quantity of rainfall, the amount of light exposure, the lifespan, the presence of pesticides, and the presence of diseases.

Since mango farmers have access to a wide range of information that is offered by these data points, they are in a position to make well-informed choices regarding the agricultural practices that they employ to cultivate mangoes, which puts them in a better position to be successful. They may, for instance, improve resource allocation by determining the ideal amounts of fertilizer, water, and pesticides required for their mango trees. This would allow them to grow more mangoes with less effort. Because of this, they would be able to

provide a higher level of care for their mango plants. As a direct consequence of this, they will not waste any of these vital resources, which enables them to save money. Because of this, agricultural practices are going to continue to develop so that they are not only more effective but also more kind to the environment that they are a part of.

Additionally, the Smart Mango platform provides farmers with the capacity to properly estimate the optimal time to harvest their mangoes, which is a key component that impacts the quantity of sweetness and taste that is contained inside the fruit itself. Mango farmers in Sri Lanka are now provided with the knowledge and skills required to optimize their resource utilization, boost their yields, and dramatically enhance their profitability as a direct result of our prediction model. This is the case even though climatic patterns and several other features of the environment are prone to change continuously.

In conclusion, Smart Mango presents a chance for a solution to the age-old issue of inaccurate mango production predictions, ushering in a new era of data-driven precision in the mango farming sector of Sri Lanka. Smart Mango was developed in Sri Lanka. much less dangerous to the environment of the neighborhood where it is located.

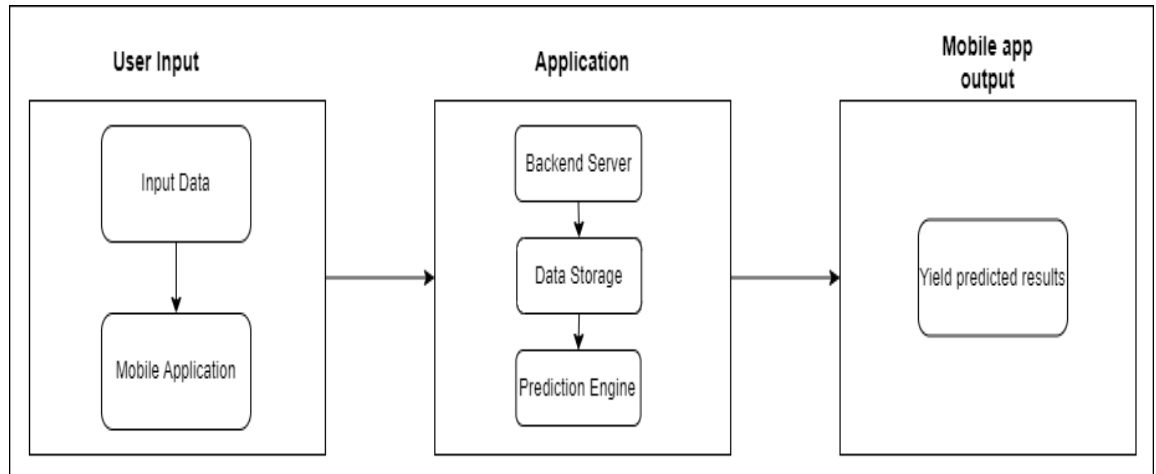


Figure 1: Application Overview

1.1 Background

This background study aims to analyze the research that has previously been undertaken on mango yield prediction, with a special emphasis on the factors that have an impact on production. In particular, the aspects that influence production will be the primary focus of this investigation. These elements include the pH of the soil, the moisture content of the soil, the temperature, the humidity, the amount of rainfall, the amount of light

exposure, the age of the mango tree, the use of pesticides, and the management of disease. If we review the literature in these areas, we will be able to get insights into the current state of mango yield prediction, identify research gaps, and highlight the possibility of future advances in this subject matter.

Soil pH: The pH or alkalinity of the soil is an important factor that determines the quality of the mangoes that may be grown there. The best circumstances for growing mango trees are soils with a pH ranging from 6 to 7.5 and conditions that range from slightly acidic to neutral. In soils with an extremely acidic or alkaline pH, the availability of nutrients may be decreased, which may have a detrimental effect not only on the health of the tree but also on the amount of product it produces. Researchers have investigated the relationship between the pH of the soil and the yield of mangoes to create ideas for practices that may be utilized to manage the soil to obtain ideal pH values. Mango yields are directly correlated with the pH of the soil.

Soil Moisture: When determining the number of mangoes that can be harvested from a given area, the amount of moisture that is available in the soil is one of the most important aspects to consider. Although mango trees need an adequate amount of moisture throughout the whole of the growth season, this demand becomes much more important during the phases of blooming and fruit development. If the soil is either too dry or too damp for the plant's habitat, there is a chance that the plant may get stressed and have fewer flowers and fewer fruits set if the plant is grown in such conditions. A rising number of companies are resorting to cutting-edge technology such as soil moisture sensors and irrigation management systems to maintain the optimal amount of soil moisture for mango production, therefore optimizing the efficiency with which water is used and ensuring that the right level of soil moisture is always present. These sorts of technological applications are becoming more and more widespread use.

Temperature: Mangoes' development as well as their production are both affected by the temperature of their surrounding environment, which is one of the most important aspects of the environment. Both the development and production of mangoes are reactions to their surrounding environment. Frost may cause damage to mango trees, which can then lead to a decrease in the quantity of fruit that the tree produces. Frost can also cause harm to other types of trees. When it comes to temperatures that are lower than freezing, mango trees are quite sensitive. On the other side, if the temperatures are very high, it may cause the plant to undergo heat stress, which would result in the fruit dropping off of the plant. More accurate yield estimations may be attained via a combination of an improved understanding of temperature trends and the implementation of preventative actions such as strategies for preventing frost. This is a possibility that can be considered plausible.

Humidity: The levels of humidity in the air influence the process by which the flowers are pollinated, the development of the fruit, and the susceptibility of the mango plant to

disease. The higher the levels of humidity in the air, the more susceptible the mango plant is to disease. There is a larger chance of insufficient pollen germination and a decrease in the pollen's viability when humidity levels are low. This danger also increases the likelihood that the pollen will not be viable. Both factors may contribute to an inadequate amount of fruit being set. On the other hand, a high humidity level may be conducive to the growth of fungal diseases such as anthracnose and powdery mildew. These conditions may be found in certain environments. Both circumstances are more likely to occur in moist surroundings. To acquire the best possible crop yields from mango orchards and to reduce the likelihood of disease outbreaks, it is necessary to adhere to demanding practices such as monitoring and maintaining the levels of humidity. To reach these objectives, it is necessary to engage in certain routines, and you should do so immediately.

Rainfall: The amount of rainfall, in addition to its pattern and location, is a crucial component that influences the number of mangoes produced. The blooming and fruiting phases of a mango tree's life cycle need a suitable amount of steady rainfall that is spread out in an even way. Mango plants need this throughout their whole life cycle. Rainfall that is either inconsistent or excessive may cause flowers to wilt, fruit to split, and an increase in the number of people who get ill. This can also cause an increase in the number of people who become sick. Supplementary irrigation systems can be of assistance in alleviating the impacts of water stress and providing more accurate output estimates in regions of the world that are exposed to erratic rainfall. These regions include several countries.

Light Exposure: The ability of the mango tree to take in light is vital for the growth of the mango tree as well as the maturation of the fruit that the tree bears. It is common knowledge that mango trees have the greatest growth potential when they are exposed to direct sunshine. This is because mango trees need a significant degree of light to produce fruit that is of good quality, which is why this is the case. If there is not enough light, there is a chance that the fruit will be smaller, the ripening process will take longer, and the yield will be lesser. All these things will be caused by the lack of light. To expand the amount of light that may enter and travel throughout the orchard, it is vital to perform appropriate pruning and canopy management practices. This may be done by allowing as much light and air into the orchard as possible.

Mango Tree Lifespan: When trying to predict future yields, one of the most significant elements to consider is the age at which mango trees will begin to produce fruit. This is because age is one of the most critical factors in determining how much fruit a mango tree will produce. Normal fruit production on mango trees begins between the ages of three and four years, and if the trees are well-cared for, they may continue to produce fruit for many decades. Mango trees typically start producing fruit between the ages of three and four years. In most cases, mango trees do not start bearing fruit until they are between the ages of three and four years old. To create an accurate projection of the amount of fruit

that will be produced by an orchard's trees, it is vital to have a solid understanding of both the age of the trees and their current state of health. Elder trees, on average, produce a bigger amount of fruit than their younger counterparts do during their lifetimes.

Pesticides: The use of the proper pesticides is an essential component of mango agriculture. This is because insects and illnesses could have a substantial effect on mango output. Fruit flies, scale insects, and aphids are the three most prevalent types of pests that attack mango trees. Mango plants are often troubled by a variety of pests. There is also the possibility that infectious diseases such as anthracnose and bacterial black spots are to blame for the loss of fruit. The use of integrated pest management (IPM) techniques not only enables a decrease in the overall amount of pesticides that are sprayed but also results in an improvement in the level of protection afforded to mango trees and the environment in the surrounding region. These techniques encompass not just biological and chemical measures, but also cultural and behavioral strategies for avoiding insect infestations.

Diseases: Diseases that might affect mango trees pose a substantial risk to the crop, both in terms of the quantity and quality of the fruit that the trees can produce. One of the most damaging mango illnesses is called anthracnose, and it is caused by a fungus called *Colletotrichum gloeosporioides*. This disease attacks mangoes and causes them to rot from the inside out. The mango fruit may be vulnerable to this disease at any stage in its development; it does not matter when it is fertilized. The plant's reduced fruit output is likely due to the presence of other diseases, such as powdery mildew, bacterial black spots, and mango scab. To effectively manage disease outbreaks and optimize production estimates, early detection, disease-resistant cultivars, and appropriate fungicide sprays are essential.

Data Analysis and Machine Learning: Recent advancements in data analysis and machine learning have made great strides in improving the accuracy of mango output forecasts. Researchers now have access to a vast treasure mine of data, which may include records of previous weather conditions, information on the soil, records of pests and diseases, and records of yields. Algorithms based on machine learning have been used to develop prediction models that simultaneously take into consideration several different elements. Regression models, decision trees, and neural networks are a few examples of algorithms that fall under this category. Mango farmers can make more informed choices because of these models' capacity to create production forecasts that are not only more accurate but also capable of being updated in real time.

Predicting mango yields is difficult because of the many factors that must be considered. These factors include the pH of the soil, the amount of moisture contained in the soil, the temperature, the humidity, the amount of rainfall, the amount of light exposure, the life span of the mango tree, the use of pesticides, and the presence of diseases. To be able to make informed decisions on mango cultivation, resource distribution, and market

planning, farmers, researchers, and policymakers need to be able to properly forecast mango yields. Because of innovations in data science, technological advancements, and agricultural practices, our ability to properly anticipate mango harvests has witnessed significant gains in recent years. All these different aspects have made a difference in the sector. Despite this, ongoing research and the ability to adapt to changing environmental conditions are essential if one is to further develop yield prediction models and ensure the sustained sustainability of mango production in the face of a changing terrain of barriers.

1.2 Literature Review

Yield prediction models that make use of machine learning algorithms have seen a boom in popularity in recent years. This is a direct result of the potential advantages that these models may provide in the context of the agricultural sector. In recent years, this trend has seen a significant increase. The potential benefits that may be provided by these models are the primary reason for their rising popularity, which may be directly attributable to this trend.

The purpose of this article is to analyze and evaluate the procedures used and the results obtained by four separate pieces of research that investigate the use of machine learning algorithms to predict agricultural yields. Specifically, this article will focus on analyzing and comparing the methods followed and the findings obtained. These investigations were carried out by various researchers at different periods and for different causes for a variety of different reasons.

In 2018, a model was presented that estimates crop yields from mango plantations by making use of artificial neural network (ANN) approaches. This was done to better serve customers. To provide accurate projections of agricultural yields, this model was constructed. A model was constructed to ensure that accurate forecasts of crop yields from mango plantations could be made. During the research, both meteorological data and information about the nutrients that are already present in the ground were utilized as input components to generate a forecast about the crop yield of mangoes. This was done to determine how many mangoes may potentially be harvested. The purpose of the activity was to calculate the total number of mangoes that could be obtained from the crop when it had matured. The findings of this study demonstrated that the ANN model could enhance mango production forecasting, and it achieved this goal with an accuracy rate of 87.4% [3].

An investigation of the several distinct approaches that may be used within the domain of machine learning was carried out in 2021 to forecast the output of agricultural companies. This was done to better understand the breadth of available options. This research was conducted to gain a deeper understanding of the variety of accessible choices. After

completing an analysis of a total of 54 distinct research articles, the researchers arrived at the conclusion that was presented above. They concluded that the machine learning algorithms that they used could be separated into a total of six distinct categories after conducting their research. This category includes a wide range of subjects, including deep learning, artificial neural networks, regression, clustering, decision trees, support vector machines, and deep learning. These are the several categories that were considered during the process of doing the research for the study. According to the results, artificial neural networks and deep learning algorithms performed better in terms of accuracy rates compared to decision tree and support vector machine approaches [4]. The task of estimating the production of tomato crops has lately been the focus of an experiment that will examine the use of machine learning algorithms in this context. Sharma and his colleagues are now carrying out this experiment. For the two researchers to get an accurate estimate of the yield of tomato crops, the inputs that they used in their model each consisted of a total of six independent features. Because of this, they were able to achieve their objective of creating an accurate estimate. The speed of the wind, the amount of precipitation that fell, the amount of solar radiation that was present, as well as the temperature of the ground all had a role in the outcome.

According to the findings, the artificial neural network methodology was better than the other techniques of machine learning since it had a higher accuracy rate than those other methods [5]. This was shown by the fact that the results. Machine learning algorithms were used by a researcher in the year 2020 to carry out an in-depth literature review of agricultural output forecast models. The researcher was responsible for carrying out this review. This inquiry was carried out by the researcher to fulfill the prerequisites for writing their dissertation. According to the conclusions of a study that looked at 109 different research articles, the algorithms that were used for predicting agricultural output most of the time were found to be artificial neural networks, support vector machines, decision trees, and regression analysis. The study was conducted to find out which algorithms were used the most often. The findings of this study indicated that the accuracy of machine-learning models is dependent not only on the kind of crop being studied but also on the input parameters [6]. This finding was reached because it was found that the accuracy of these models varies depending on the input parameters.

In conclusion, the results of the research study that were discussed in this literature review show that it may be feasible to improve yield prediction for several crops by using tactics that are related to machine learning, in particular artificial neural networks. These findings were discussed in this article because they were included in the discussion of the research study that was reviewed in this article. These results were provided within the context of a discussion of previous research that had been conducted on the subject. Following the compilation of these facts into a hypothesis, more research was carried out to put this idea to the test. On the other hand, the accuracy of these models may be impacted not only by

the characteristics of the components that are being input into the model but also by the crop that is the major focus of the inquiry. This is because the crop is a more complex system than the individual components. It is necessary to research to build yield prediction models that are dependable and accurate for a greater variety of crop kinds and growing conditions. Only by doing more research can this goal be attainable. Regarding this subject, more research is necessary.

1.3 Research Gap

Although the literature study included a vast array of studies that made use of machine learning techniques to create yield prediction models for mango crops, there is still a dearth of research in this industry area. Even if the literature review was carried out, the situation is still the same as it was before. There has not been nearly enough study done that specifically investigates how various environmental and soil parameters impact prediction models for mango production.

The research that was conducted by M. R. Islam and colleagues (2018) examined the production of mangoes by making use of data about the climate and the soil. Although they were successful in this quest, they did not perform an investigation of the influence that certain climatic and soil parameters had on the forecast of yields.

It is not completely implausible to think that in the not-too-distant future, researchers will focus their attention on determining the impact that a variety of variables have on mango yield prediction models. Some of these characteristics include the temperature, the quantity of precipitation, the moisture levels in the soil, and the amount of nutrients that are present. Inadequate research has also been conducted in other areas, such as the use of several distinct machine-learning algorithms to predict mango output. On the other hand, a variety of alternative techniques to machine learning, including support vector regression and random forests, are effective in the process of predicting agricultural outputs. In the research that was carried out by M. R. Islam and colleagues (2018), an artificial neural network model was employed to make forecasts about mango production. The following study [7] examines how well several machine learning algorithms predict future output and compares the findings of each of these algorithms. Mango farmers who cultivate mangoes may have something to gain from this research. This research may be of service to them in selecting the algorithm that is best for their operations both in terms of accuracy and efficiency, and that is one of the reasons why this study was carried out in the first place.

Although there has been some research conducted on the use of machine learning models to forecast mango crop output, probably, further research on the many different kinds of mango is still required. Certain types of mango have received very little attention in terms

of scientific investigation. There is a good chance that various varieties of mangoes have varying requirements when it comes to the sort of soil, the degree of fertilization, and the level of defense against diseases and pests. Because of this, the influence that each of these unique elements has on productivity may likewise vary depending on the circumstances. Testing and validation within a constrained range of possible applications: To ensure that machine learning models are accurate and trustworthy, these models need to be validated and examined by making use of the data that was collected. Utilization of the data is the only way to achieve this goal. It is not only possible for models that have been trained using machine learning to accurately estimate yield, but it is also quite important that these models be able to do so. There is a possibility that more research is required to test and evaluate machine learning algorithms for agricultural output prediction [8]. This would be done with the aim of testing and assessing the algorithms. It is quite likely that this will need more inquiry into the many kinds of mangoes that are available.

Table 1: Research Gap

Study	ML Technique	Input Factor	Output	Accuracy
Islam et al. (2018)	Artificial neural network	Temperature, rainfall, humidity, soil moisture	Mango yield	Medium
Garg et al. (2021)	Regression, decision tree, neural networks, ensemble methods	Diverse input factors across multiple crops	Crop yield	Low
Sharma et al. (2020)	Super vector regression, decision trees, random forests, artificial neural networks	Temperature, humidity, soil moisture	Tomato yield	Medium
Berghout et al. (2020)	Diverse ML techniques across multiple crops	Environmental factors	Crop yield	Low
Smart Mango	Regression decision tree model incorporating pesticide and diseases	Soil pH, Soil Moisture, temperature, humidity, Rainfall, Light Exposure, Life Span, pesticides, and diseases	Mango yield	High

2. RESEARCH PROBLEM

The problem that this study intends to address as a concern in the research community is the absence of accurate and reliable yield prediction models for mango agriculture in Sri Lanka. This is the issue that this study aims to solve. This research aims to find a solution to this issue. Finding a resolution to this specific issue is the purpose of the work that is being done here. Although mango growing is a significant industry in the country, it is not easy to make an accurate prediction of the number of mangoes that will be produced. Even though mango output is notoriously difficult to forecast, this is nonetheless the case. However, even though the cultivation of mangoes is of utmost significance, this is the current situation. Not only does the unpredictability of mango harvests have a direct impact on the amount of profitability that may be achieved by mango farmers, but it also has a direct influence on the choices that mango farmers can make as a consequence of their financial condition. In other words, the unpredictability of mango harvests has a direct influence on the amount of profitability that may be gained by mango farmers.

Most of the recent work that has been done to estimate mango production has focused on making use of techniques from the field of machine learning to develop models that are based on a wide range of environmental and soil factors. This has been the primary focus of most of the recent work that has been done. This has been the focus of most of the work that has been done within the most recent few years. This has been the focus of the investigation's bulk of its efforts throughout its whole. These models were designed in the first place so that estimates of mango output could be determined from the data that they supply. Mango production estimates may be calculated from these models. These factors include but are not limited to, the temperature, the humidity, and the amount of moisture that is contained within the soil. Additionally, the quantity of moisture that is contained within the soil is also a factor. In addition to this, another factor to consider is the amount of moisture that is already present in the soil. Another factor that must be considered is the quantity of moisture that is already contained within the soil. Despite the findings of several studies that suggest there is a possibility that yield estimations might be improved, there is presently a dearth of research being conducted in several crucial areas, including the following:

Because of the paucity of resources, the research on the many distinct varieties of mangoes was limited, and consequently, the findings were not as good as they might have been. Because there has not been a great lot of research done in this field to analyze the unique requirements and qualities of the many types of mango that are accessible, there is a lack of information on this particular topic as a consequence of the fact that there is a lack of knowledge in this particular subject. This is the key factor that contributes to the restricted quantity of information that can be retrieved at any one time. Different varieties of mango might respond in a variety of one-of-a-kind ways based on the conditions of the soil and the environment that surrounds them. This is something that should be considered.

Because of this, many models may need to be constructed to properly consider the different kinds of mango.

The act of putting an assertion or hypothesis to the test and verifying the results. There is an urgent need for severe validation and testing to assess the accuracy and reliability of these models across a range of mango farms and locations in Sri Lanka. This can only be done by conducting extensive tests. Although machine learning models have been proposed for yield prediction, there is a pressing need for testing and validation to take place in an extremely timely manner. Because of this, it will be feasible to get a more in-depth knowledge of the capabilities of machine learning models to make yield predictions.

It Is Necessary to Take into Consideration Every Aspect That Is Linked to the Relationship That Exists Between Pesticides and Illnesses. The "Smart Mango" model that has been built knows to properly anticipate production levels, it is important to consider elements related to pesticides and diseases. This is something that has been taken into consideration by the model. This was done to make the model's portrayal of reality more realistic, and it was done so by doing this. Ultimately, the success in reaching this objective was attributable to the fact that the "Smart Mango" concept was kept in mind. It was important to develop a model that makes use of several different components of artificial intelligence to accomplish this objective most effectively. However, further research is necessary to get an understanding of the extent to which the aforementioned factors have an impact on mango yields and the several ways in which these aspects may be effectively included in forecasting models. Both the yields of mangoes and the plethora of diverse ways in which the criteria may be successfully included in prediction models are subjects that require more study. Specifically, the yields of mangoes are a field that needs more investigation. On each of these subjects, there is a need for further research to be conducted.

In the field of machine learning, the practice of using a broad variety of learning approaches in a variety of different configurations. In the context of mango yield prediction, it is of the utmost importance to compare and evaluate the effectiveness of these various methodologies. Previous research used a wide variety of distinct strategies for the process of machine learning. Despite this, it is of the utmost importance to compare and evaluate the effectiveness of a variety of methodologies in terms of their capacity to anticipate mango output. Farmers are looking for guidance in selecting the algorithm that would perfectly and effectively accommodate the activities that are unique to their business, and they expect to get this advice from the Internet. Farmers are looking for direction in selecting the algorithm that would exactly and effectively accommodate the activities that are unique to their company. Farmers are seeking guidance in picking the algorithm that can exactly and effectively handle the tasks that are unique to their line of work.

In a word, the purpose of the project is to develop reliable and precise models for predicting mango yields, which will serve as the project's deliverable. These models are supposed to take into account the extensive variety of mango varieties; they are also supposed to include factors linked to a variety of illnesses and pests that are associated with mangoes; they are supposed to evaluate the performance of the model; and they are supposed to compare and contrast several different machine learning methodologies. The resolution to this problem will be of significant benefit to farmers who cultivate mangoes in Sri Lanka since it will make it possible for those farmers to have access to data-driven insights that will assist them in improving the agricultural practices that they now use. As a result, the solution will be of significant value to mango farmers in Sri Lanka. Because it will make it feasible for mango farmers in Sri Lanka to have access to data, this resolution will be of significant advantage to mango farmers in Sri Lanka. Mango farmers in Sri Lanka will profit significantly from this resolution. As a result of this, not only will there be an increase in the total quantity of output, but there will also be an increase in the amount of money earned.

3. RESEARCH OBJECTIVE

3.1 Main Objective

It is important to take into account factors such as soil pH, soil moisture, temperature, humidity, rainfall, light exposure, life span, pesticides, and disease infestations when developing and improving machine learning models for accurate and reliable yield prediction of particular mango variants with limited and varied data availability.

3.2 Sub-objective 1:

collecting the necessary information from agricultural enterprises and research institutes situated in the surrounding region to uncover the key environmental and agronomic factors that affect the yield of different mango varieties. To examine and contrast several different machine learning strategies for yield prediction, such as regression, decision trees, neural networks, and ensemble methods; to select the most effective technique based on performance measures such as memory, accuracy, and precision; and to provide findings and conclusions after completing these tasks.

3.3 Sub-objective 2:

Build and train machine learning models by using the algorithm that has been supplied, and the data that has been gathered, taking into consideration the particular needs and characteristics of the various mango varieties, as well as the meteorological conditions that are unique to the region, and taking into account the data that has been acquired.

3.4 Sub-objective 3:

assessing the accuracy and dependability of the yield projections made by the models, as well as validating and testing the models that have been created with the assistance of several different data sets. to investigate a variety of techniques for enhancing performance, such as model assembly, feature engineering, and hyperparameter tweaking, and then to tailor the optimization of the models to the results of the validation after those results have been compiled.

3.5 Sub-objective 4:

The purpose of this project is to develop a mobile application that is straightforward to use and provides farmers with the opportunity to enter the necessary data to get yield estimates for various mango types. Additionally, farmers will be provided with guidance to improve their agricultural practices based on the output that is anticipated by the program.

4. METHODOLOGY

An essential component of this investigation is the research approach that was used to arrive at an informed estimate of the quantity of mangoes that will be produced in Sri Lanka. This was accomplished by conducting a survey. The project will employ a systematic approach that will include data gathering, model construction, validation, and evaluation to solve the issue of erroneous mango yield predictions. This will be done to address the problem. To discover a solution to the problem, this step will be taken. The methodology may be broken down into several essential components, each of which plays an important part in accomplishing the research goals that were given at the beginning of the study. The research objectives were defined at the beginning of the investigation.

Data Collection: The first thing that must be done is to gather relevant data from a broad range of sources. This is a very necessary step. This document provides information on a variety of mango varieties, as well as historical data on mango production, data on weather and soil conditions, data on pests and diseases, and historical data on the yield of mangoes. The prospective data sources might include mango farms, government agricultural departments, meteorological agencies, research institutes, and other academic organizations.

The Data Are Being Preprocessed After the data have been obtained, they will need to be sorted, cleaned, and preprocessed so that they are of a high quality and consistent across the whole set. To resolve this issue, you will need to address issues with the formatting, as well as values that are missing and outliers. To ensure that the prediction models provide accurate results, it is necessary to preprocess the data.

Feature Selection and Engineering: Feature Selection: Once the data have been preprocessed, the next step is to choose the features or variables that are going to be used in the prediction models that are going to be the most pertinent. This stage comes after the step of deciding which features or variables are going to be used. To do this, it is necessary to identify the factors that have the most impact on mango production. These factors include but are not limited to, temperature, humidity, soil moisture, and the presence of pests and diseases.

Engineering New Features: Depending on the specifics of the situation, it may be possible to create new features by drawing on existing expertise in a particular field. For instance, to identify the degree of environmental stress that mango trees are experiencing, one may calculate a "stress index" by integrating data on temperature, humidity, and rainfall. This would allow for the determination of the level of environmental stress that mango trees are experiencing.

Model Development: Methods available for selection. To make an accurate forecast of mango production, we will be using a variety of methods, including statistical modeling

and machine learning. Typical examples of algorithms include linear regression, decision trees, random forests, support vector machines, and neural networks. Other kinds of algorithms include random forests and support vector machines. The selection of relevant algorithms will be directed by the qualities of the data that are going to be studied as well as the objectives of the research.

Model Training: To educate the algorithms that were selected, a subset of the collected data will be utilized, and this subset is known as the training set. The models gain a grasp of the relationships that exist between the mango yield objective variable and the input variables, which are sometimes referred to as features, while they are being trained. This understanding is essential for the models to have. It is possible to utilize techniques of cross-validation to assess the performance of a model while it is still in the process of being trained.

Validation Dataset: To examine the models' ability to generalize, a unique dataset (the validation set) that the models have never seen before will be used for validation. This will be done so that the models' capacity for generalization may be evaluated. This helps in the identification of overfitting, which is when the models perform well on the data that they were trained on but poorly on new data. In this scenario, the models perform well on the data that they were trained on.

Evaluation Metrics: To assess the effectiveness of the model, several different evaluation metrics will each play a part in the process. The values for Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R²) are a few examples of the metrics that may be included in this category. When deciding the measurements to utilize, the specific objectives of the research will be a significant factor to consider.

Model Fine-Tuning and Optimization: Adjustments to the Model's Hyperparameters Needed This technique, which is referred to as "hyperparameter tuning," is carried out to fine-tune the hyperparameters of the models (such as learning rate, tree depth, and regularization strength). It is feasible to locate what you want by using search strategies such as grid search or random search.

Techniques for Forming Ensembles It is feasible to aggregate the predictions of many models to improve accuracy by using ensemble techniques such as stacking or boosting, for example. This may be done to make use of multiple models' outputs.

Model Interpretability: The ability to explain is of the highest significance since it is essential to understand how the prediction models arrive at their findings. The implementation of interpretability methodologies such as feature significance analysis and model visualization will be required to explain the components that go into mango yield forecasts.

Testing and Deployment: To determine how effectively the final prediction models, work in the real world, they will each be assessed using a distinct test dataset. This will allow us to determine how well they perform.

After the models have been validated via testing and demonstrated to provide reliable results, they may then be deployed as a component of a web-based or mobile application. Growers of mango in Sri Lanka would be able to acquire production predictions that are customized to their farm conditions because of this development.

Continuous Monitoring and Updating: Once the predictive models have been implemented, they should be constantly examined and updated whenever there is new data available. This is to ensure that the models continue to accurately anticipate the future. This ensures that the models will, over time, continue to be accurate and relevant in their representation of reality.

Communication and Outreach: The results of the research, including its findings and conclusions, will be presented to mango farmers, agricultural groups, and government officials in Sri Lanka. Building instructional materials, workshops, and seminars may be something that could be done to make it easier for people to talk to one another about the use of predictive models and the benefits that they provide.

Ethical Considerations: Data Privacy The confidentiality of your information as well as the preservation of your privacy will be of the utmost importance. Before being handled in a method that is by the ethical requirements of the research, each piece of data that is used in the investigation will first be turned anonymous.

Limitations and Future Investigation: The investigation will acknowledge that it is constrained in several ways, including the potential for bias in the data sources that were utilized and the challenge of accurately estimating all of the factors that have the potential to have an impact on mango output.

Future Research: Recommendations for future research pathways may be offered, which may include exploring the use of artificial intelligence algorithms that are more powerful and integrating data that is more fine-grained.

In conclusion, the use of this method provides a methodical approach to addressing the problem of inaccurate mango production estimates in Sri Lanka. Since we first became aware of this issue, it has been a source of worry. This project's goals are to provide substantial new insights and tools to mango farmers in the region, with the end goal of improving mango farming practices and increasing the farmers' overall profitability. The gathering of data, followed by its preprocessing, the creation of prediction models, and the verification of the models' correctness will be how these objectives will be met.

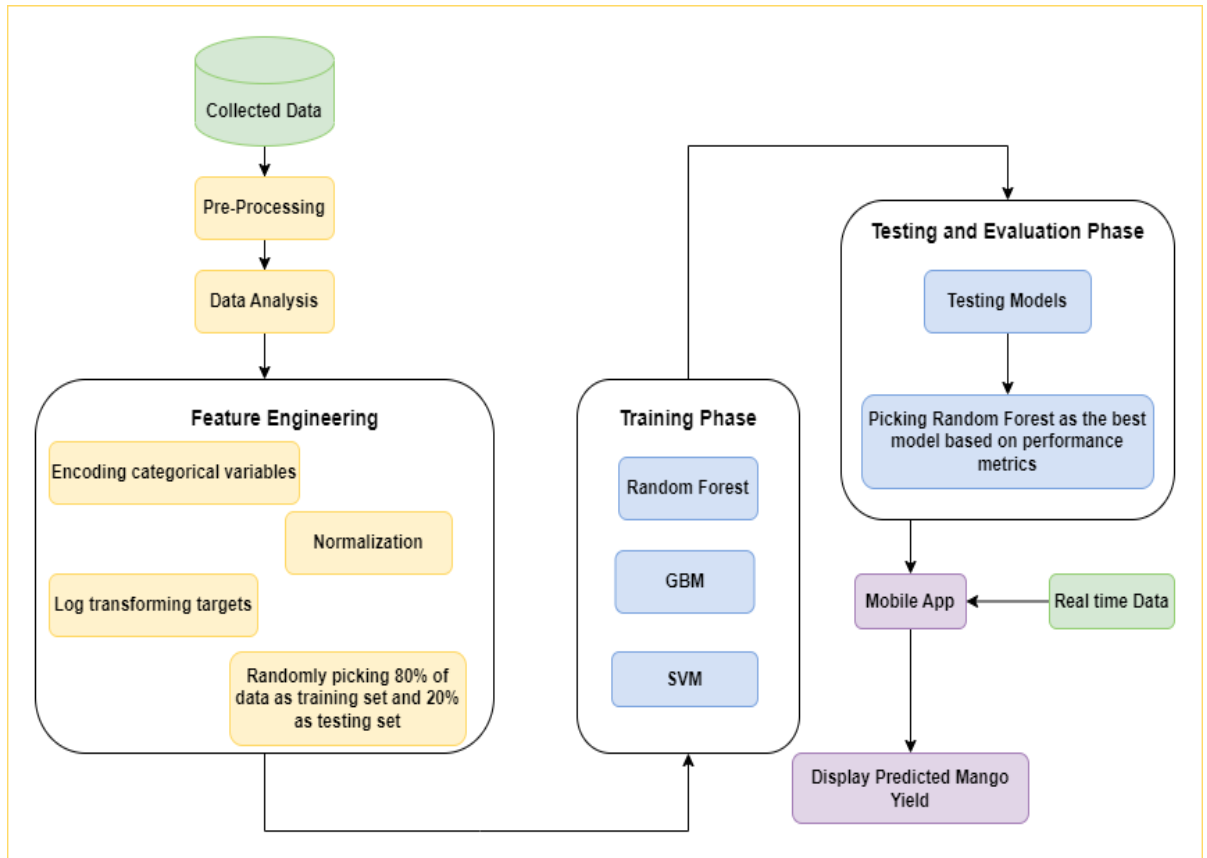


Figure 2: Overall Architecture

The above diagram illustrates data Collected for Input This section is where you will enter any data you have collected about mango farming. This data might contain information about the meteorological conditions, the qualities of the soil, how much fertilizer was used, how much irrigation was done, and data on past yields.

	Soil pH	Soil Moisture	Temperature	Humidity	Rainfall	Light Exposure	Life Span	Yield	pesticide_1	pesticide_2	Disease_1	Disease_2
0	5.592830	0.472886	37.896900	49.677238	1515.167403	9.661759	76.480229	456.540104	0	0	0	0
1	4.722204	0.280567	39.295862	30.317420	477.582412	2.978146	15.213885	177.678835	1	0	1	1
2	5.761180	0.090315	8.099947	18.305295	2308.356148	10.674392	95.878379	311.165335	0	1	0	1
3	6.124865	0.510815	15.428293	80.393577	2180.486784	18.090881	62.560397	938.603124	0	0	0	0
4	7.433218	0.672366	14.099188	55.453696	1760.748061	7.972518	95.674312	649.156705	0	0	0	0
...
99995	5.991148	0.292031	21.414808	43.652343	1428.919659	11.666642	86.654768	482.642006	0	0	0	0
99996	6.326121	0.120369	27.336862	68.384733	1195.797977	7.185621	32.831322	435.882376	0	0	0	0
99997	8.423406	0.749855	25.172054	84.223089	683.247433	2.396103	58.246464	537.266930	0	0	0	0
99998	8.919418	0.924381	12.574419	69.805588	646.663498	6.517845	22.361164	773.637808	0	0	0	0
99999	7.983554	0.745405	9.738876	65.494734	1662.879438	10.306439	42.819035	783.930429	0	0	0	0

100000 rows x 12 columns
Warning: total number of rows (100000) exceeds max_rows (20000). Limiting to first (20000) rows.
Warning: total number of rows (100000) exceeds max_rows (20000). Limiting to first (20000) rows.

Figure 3: Data set

Before feeding the data into machine learning models, it is typically necessary to clean and alter it first. This step is referred to as preprocessing. The handling of missing data, the elimination of outliers, and the conversion of data into a format that is appropriate for analysis are all examples of preprocessing steps.

The next step, "Data Analysis," entails conducting exploratory data analysis to get insights into the dataset. You will have a better understanding of the interactions between the various variables and the ability to recognize patterns that may be significant when it comes to estimating mango yield.

Building predictive models requires numerous steps, one of the most important of which is feature engineering. Creating new features from the existing data or altering existing features to more accurately depict the relationships contained within the data are both required steps in this process. Encoding categorical variables, performing normalization, and log-transforming targets (which may be yield values) are all discussed in this instance.

A training set and a testing set are separated from one another within the dataset, which has been split into two sections. While machine learning models are trained on the training set, which consists of 80% of the data, the performance of the model is later evaluated using the testing set, which consists of 20% of the data and is kept separate.

Training Phase: This stage of the process involves applying several machine-learning algorithms to the training data. Random Forest, Gradient Boosting Machine (GBM), and Support Vector Machine (SVM) are the three algorithms that are discussed in the graphic. Each of these algorithms is trained on the training data to understand patterns and correlations between the features and the target variable, which in this case is the yield of mangoes.

Training and Evaluation: Once the models have been trained, the testing data is used to conduct the evaluation. To evaluate the efficacy of each model's ability to forecast mango

production, performance indicators such as accuracy, mean squared error, and others are utilized. According to these measures, the diagram indicates that the Random Forest model should be selected as the one that performs the best.

Mobile App (Real-time Data Input): During this stage of the process, the selected model, which is the Random Forest model in this instance, is implemented into a mobile app. The real-time data input takes place. The application can take inputs of real-time data, such as the present weather conditions, the levels of soil moisture, and any other relevant parameters that affect mango yield.

Predict Mango Production The system can make predictions about mango production by employing a trained Random Forest model and real-time data inputs from the mobile app. The model makes yield forecasts by making use of the data that is inputted and the patterns that it has learned.

```
synthetic_data = pd.DataFrame({
    'Soil pH': 6.5,
    'Soil Moisture': 0.4,
    'Temperature': 25.0,
    'Humidity': 60.0,
    'Rainfall': 1000.0,
    'Light Exposure': 8.0,
    'Life Span': [80.0],
    "pesticide_1" : 0 ,
    "pesticide_2" :1 ,
    "Disease_1" : 1 ,
    "Disease_2" : 1
})

# # Make predictions on the synthetic data row
predictions = model.predict(synthetic_data)

# Print the predicted yield
print('Predicted Yield:', predictions[0])
```


 Predicted Yield: 530.0397248838013

Figure 4: Synthetic data row

This figure provides an overall illustration of a data-driven strategy for estimating mango yield. In this strategy, machine learning models are trained on historical data and then implemented in a mobile app. The app then provides real-time yield estimates based on the current environmental and agricultural circumstances.

4.1 Testing and Implementation

This system is tested throughout the development lifecycle. All the modules were tested separately before and after the integration with the main system. The testing of the entire system was done to ensure the proper functioning of the system.

Table 2: Testing results

Test ID	Test Description	Test Results
T001	Testing of the ML model by varying the input values	Pass
T002	Testing of the IoT module with the ML module	Pass
T003	Testing of the Database	Pass
T004	Testing of the mobile application	Pass

5. RESULTS AND DISCUSSION

5.1 Results

This chapter presents experiment results in the results section and the findings of those experiments in the research findings section. Finally, the discussion section summarizes the findings and the reasoning behind these findings.

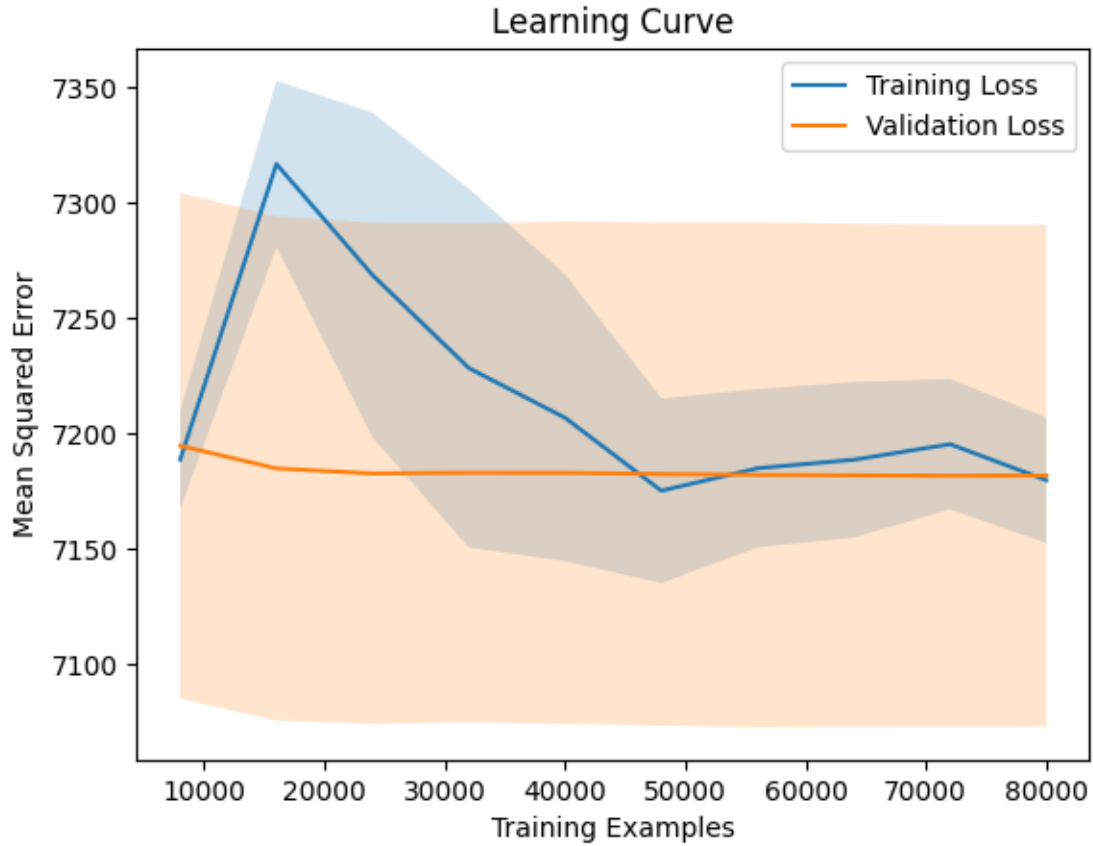


Figure 5: Learning Curve

Figure 5 depicts how the performance of the constructed ML model changes as the number of training instances (X-axis) rises, often in terms of MSE or some other evaluation metric (Y-axis). This figure may be seen above. It helps you evaluate how effectively your model is learning from the data and if it is overfitting or underfitting the data.

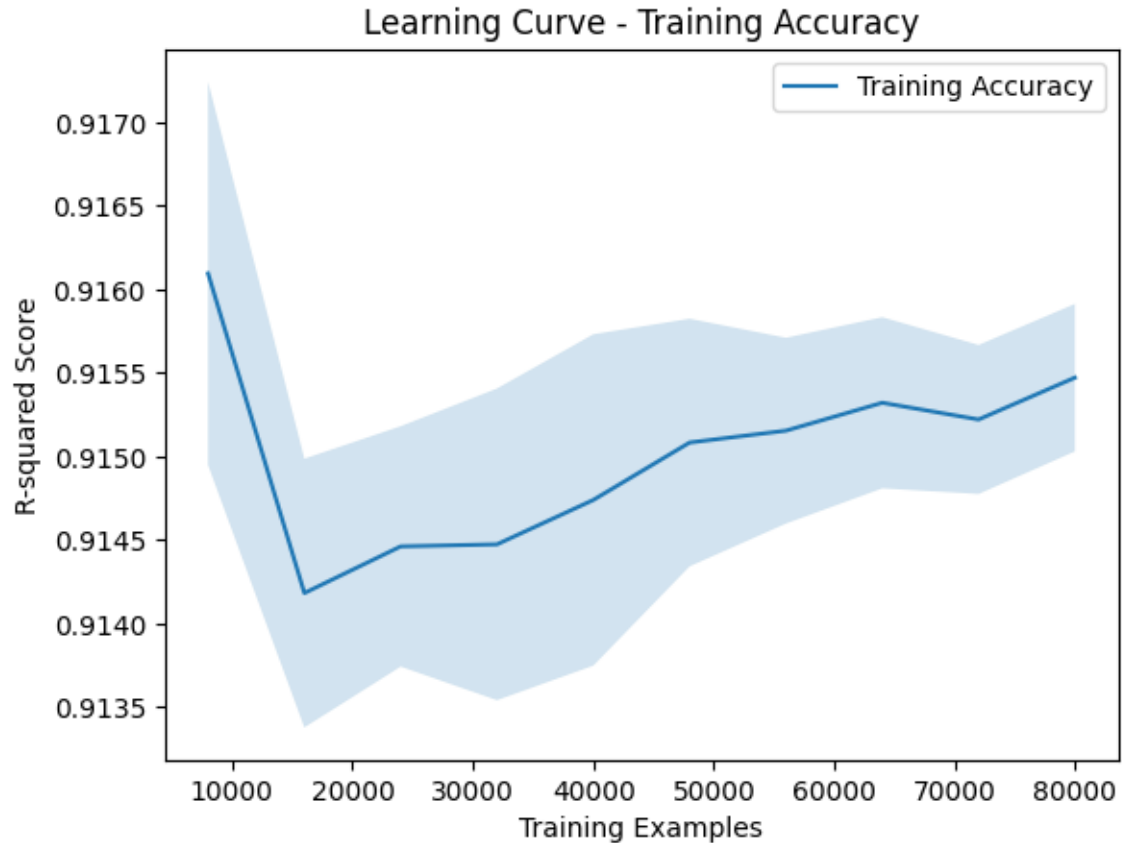


Figure 6: Training Accuracy

The R-squared score is a statistical metric that is used to evaluate how well a regression model fits the data, and it is shown along the Y-axis of the graph. It indicates the degree to which the model's predictions correspond to the actual data points. A score of 0 for R-squared implies that the model does not explain any of the variability in the data, while a score of 1 shows that the model fits the data perfectly.

The number of training instances that were used to train the machine learning model is shown along the X-axis. This axis illustrates how the performance of the model shifts in response to changes in the total quantity of data that was used for training. In most cases, it will begin with a limited number of samples and then steadily develop from there.

The value of R squared comes to 91%. A score of 91% is high when considered in the perspective of R squared. It is an indication that your model can explain a large percentage of the variability seen in the data and that it can produce relatively reliable predictions. It seems as if your model is doing an excellent job of describing the variability in the data and producing accurate predictions based on those facts.

5.2 Discussion

The purpose of the research that was carried out for this study was to forecast mango production using a Random Forest machine learning model, taking into consideration a variety of significant characteristics. This research was carried out to fulfill the purpose of this study. Within the scope of this study, the findings of this investigation will be discussed. This research was carried out so that the goals of this study may be met by the time the study is finished. Because they are pertinent to the scope of this research, the outcomes of this investigation will be investigated as part of it. This research was carried out with the hope that the goals of the study would, at the very least, be satisfied by the time the investigation is over. Because these results are pertinent to the overarching purpose of this research project, they will be investigated as a component of this investigation as part of this research. It was discovered that the model had an astounding accuracy of 91%, which is proof of the model's outstanding talents in terms of its capacity to foresee the future. This potential to anticipate the future is shown by the fact that the model had an accuracy of 91%. This capacity was shown by the fact that it had an extremely high rate of success in its operations. During this discussion, in addition to the results that are most important in and of themselves, we are going to assess the discoveries made during this study that have the most significant consequences for the future.

According to the results of this study, using machine learning, and in particular the Random Forest model, is an efficient method for predicting mango yields based on a varied range of parameters that are relevant to the matter at hand. In particular, it was shown that the model is highly effective in accomplishing the goal that it was designed to achieve. The data that was acquired from a significant quantity of mango plants made it feasible for this to be identified. Mango plants were used as the source of information. In conclusion, the outcomes of this research provide light on the possible advantages that may be obtained via the use of machine learning. The fact that such a high degree of accuracy was achieved is rather fascinating to learn about since it shows that there are new opportunities for the development of better mango farming practices and crop management. It will be required to conduct more research and put the findings of this study into practice to acquire an in-depth grasp of the possible advantages that this study may be able to supply to the agricultural industry. This study may be able to provide the agricultural industry with a greater understanding of how to improve crop yields. After that, it will be feasible to gain a full grasp of the prospective benefits that this research may be able to give to the agricultural organization. This comprehension will not be achievable till then.

6. CONCLUSION

In this study, we attempted to estimate mango yields by using a machine learning model called Random Forest. This model considered several different variables, such as soil pH and moisture, temperature and humidity, rainfall, and light exposure, the length of the mango plant's life, various pesticides and illnesses, and the plant's life span. The model achieved an impressive level of accuracy, coming in at 91% overall. The results of this research shed light on several significant implications and discoveries.

To begin, the incorporation of a wide range of variables into the forecasting model demonstrates how complicated the process of establishing mango production is. The characteristics of the soil, the circumstances of the environment, the prevention and treatment of pests and diseases, and the length of time the crop can mature are all important factors in mango production. Mango growers and the agricultural business may considerably benefit, both individually and collectively, from a better understanding and use of these aspects.

The use of the Random Forest model proved to be successful in accurately representing the intricate connections between the many input factors and the amount of mango produced. However, to guarantee the model's dependability when applied to real-world circumstances, it is necessary to verify the performance of the model by using suitable methods such as cross-validation.

When it comes to practical applications, it is necessary to determine which elements have the greatest predictive relevance. This information may help farmers make more educated choices about irrigation, fertilization, and pest management, which will eventually lead to improved crop yields and more efficient use of resources.

The successful use of this prediction model paves the way for intriguing new avenues of investigation in the future. Researchers could test more sophisticated machine learning methods, include more comprehensive datasets, and study the ramifications for both the economy and the environment. The use of sensitivity analysis and the use of technology that are used in precision agriculture might further improve mango farming practices.

In conclusion, this study demonstrates the potential of machine learning to improve mango farming by properly forecasting crop yields. Mango farming is an important part of the global mango industry. The accuracy of the model, which is 91%, is encouraging, and it opens the path for data-driven decision-making in the agricultural industry. We may contribute to the development of a more sustainable and effective method of mango production if we continue to progress this study and incorporate its findings into actual farming practices. This will be to the advantage of both mango producers and mango consumers.

References

- [1] M. Roelfsema *et al.*, “Taking stock of national climate policies to evaluate implementation of the Paris Agreement,” *Nature Communications* 2020 11:1, vol. 11, no. 1, pp. 1–12, Apr. 2020, doi: 10.1038/s41467-020-15414-6.
- [2] C. F. Schleussner *et al.*, “Science and policy characteristics of the Paris Agreement temperature goal,” *Nature Climate Change* 2016 6:9, vol. 6, no. 9, pp. 827–835, Jul. 2016, doi: 10.1038/nclimate3096.
- [3] “SRI LANKA UPDATED NATIONALLY DETERMINED CONTRIBUTIONS”.
- [4] B. Doda, C. Gennaioli, A. Gouldson, D. Grover, and R. Sullivan, “Are Corporate Carbon Management Practices Reducing Corporate Carbon Emissions?,” *Corp Soc Responsib Environ Manag*, vol. 23, no. 5, pp. 257–270, Sep. 2016, doi: 10.1002/CSR.1369.
- [5] E. P. Olaguer, “Emission Inventories,” *Atmospheric Impacts of the Oil and Gas Industry*, pp. 67–77, Jan. 2017, doi: 10.1016/B978-0-12-801883-5.00007-3.
- [6] N. R. A. Rahman, S. Z. A. Rasid, and R. Basiruddin, “Exploring the Relationship between Carbon Performance, Carbon Reporting and Firm Performance: A Conceptual Paper,” *Procedia Soc Behav Sci*, vol. 164, pp. 118–125, Dec. 2014, doi: 10.1016/J.SBSPRO.2014.11.059.
- [7] S. J. Arceivala, *Opportunities in Control of Carbon Emissions and Accumulation*. McGraw-Hill Education, 2014. Accessed: Jul. 17, 2022. [Online]. Available: <https://www.accessengineeringlibrary.com/content/book/9781259063732/chapter/chapter3>
- [8] T. Gao, Q. Liu, and J. Wang, “A comparative study of carbon footprint and assessment standards,” *International Journal of Low-Carbon Technologies*, vol. 9, no. 3, pp. 237–243, Sep. 2014, doi: 10.1093/IJLCT/CTT041.

- [9] J. Downie and W. Stubbs, “Corporate Carbon Strategies and Greenhouse Gas Emission Assessments: The Implications of Scope 3 Emission Factor Selection,” *Bus Strategy Environ*, vol. 21, no. 6, pp. 412–422, Sep. 2012, doi: 10.1002/BSE.1734.
- [10] C. C. Spork, A. Chavez, X. G. Durany, M. K. Patel, and G. V. Méndez, “Increasing Precision in Greenhouse Gas Accounting Using Real-Time Emission Factors,” *J Ind Ecol*, vol. 19, no. 3, pp. 380–390, Jun. 2015, doi: 10.1111/JIEC.12193.
- [11] J. Frijns, “Towards a common carbon footprint assessment methodology for the water sector,” *Water and Environment Journal*, vol. 26, no. 1, pp. 63–69, Mar. 2012, doi: 10.1111/J.1747-6593.2011.00264.X.
- [12] V. Franco, M. Kousoulidou, M. Muntean, L. Ntziachristos, S. Hausberger, and P. Dilara, “Road vehicle emission factors development: A review,” *Atmos Environ*, vol. 70, pp. 84–97, May 2013, doi: 10.1016/J.ATMOENV.2013.01.006.
- [13] N. P. Cheremisinoff, “Pollution Management and Responsible Care,” *Waste*, pp. 487–502, Jan. 2011, doi: 10.1016/B978-0-12-381475-3.10031-2.
- [14] G. Bekaroo, D. Roopowa, and C. Bokhoree, “Mobile-Based Carbon Footprint Calculation: Insights from a Usability Study,” *2nd International Conference on Next Generation Computing Applications 2019, NextComp 2019 - Proceedings*, Sep. 2019, doi: 10.1109/NEXTCOMP.2019.8883622.
- [15] A. C. Peres Vieira, E. M. F. da Silva, and V. V. V. Aguiar Odakura, “Development of a Web Application for Individual Carbon Footprint Calculation,” *Proceedings - 2021 47th Latin American Computing Conference, CLEI 2021*, 2021, doi: 10.1109/CLEI53233.2021.9640099.
- [16] D. Andersson, “A novel approach to calculate individuals’ carbon footprints using financial transaction data – App development and design,” *J Clean Prod*, vol. 256, p. 120396, May 2020, doi: 10.1016/J.JCLEPRO.2020.120396.

- [17] X. Yang, D. Lo, X. Xia, L. Bao, and J. Sun, “Combining Word Embedding with Information Retrieval to Recommend Similar Bug Reports,” *Proceedings - International Symposium on Software Reliability Engineering, ISSRE*, pp. 127–137, Dec. 2016, doi: 10.1109/ISSRE.2016.33.
- [18] J. Mulrow, K. Machaj, J. Deanes, and S. Derrible, “The state of carbon footprint calculators: An evaluation of calculator design and user interaction features,” *Sustain Prod Consum*, vol. 18, pp. 33–40, Apr. 2019, doi: 10.1016/J.SPC.2018.12.001.
- [19] D. Hu *et al.*, “Recommending Similar Bug Reports: A Novel Approach Using Document Embedding Model,” *Proceedings - Asia-Pacific Software Engineering Conference, APSEC*, vol. 2018-December, pp. 725–726, Jul. 2018, doi: 10.1109/APSEC.2018.00108.
- [20] Y. Wang *et al.*, “A comparison of word embeddings for the biomedical natural language processing,” *J Biomed Inform*, vol. 87, pp. 12–20, Nov. 2018, doi: 10.1016/J.JBI.2018.09.008.
- [21] Defra, “Environmental Reporting Guidelines,” 2019, Accessed: Jan. 23, 2022. [Online]. Available: www.nationalarchives.gov.uk/doc/open-government-licence/
- [22] M. Bhavadharani, M. P. Ramkumar, and S. G. S. R. Emil, “Performance analysis of ranking models in information retrieval,” *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, pp. 1207–1211, Apr. 2019, doi: 10.1109/ICOEI.2019.8862785.
- [23] L. Xiaoli, Y. Xiaokai, and L. Kan, “An improved model of document retrieval efficiency based on information theory,” *J Phys Conf Ser*, vol. 1848, no. 1, p. 012094, Apr. 2021, doi: 10.1088/1742-6596/1848/1/012094.
- [24] R. Ali *et al.*, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents Text Mining,” *Article in International Journal of Computer Applications*, vol. 181, no. 1, pp. 975–8887, 2018, doi: 10.5120/ijca2018917395.

- [25] M. Farouk, “Measuring text similarity based on structure and word embedding,” *Cogn Syst Res*, vol. 63, pp. 1–10, Oct. 2020, doi: 10.1016/J.COGSYS.2020.04.002.
- [26] B. Wang, A. Wang, F. Chen, Y. Wang, and C. C. J. Kuo, “Evaluating word embedding models: methods and experimental results,” *APSIPA Trans Signal Inf Process*, vol. 8, pp. 1–14, 2019, doi: 10.1017/ATSIP.2019.12.
- [27] S. Lai, K. Liu, S. He, and J. Zhao, “How to generate a good word embedding,” *IEEE Intell Syst*, vol. 31, no. 6, pp. 5–14, Nov. 2016, doi: 10.1109/MIS.2016.45.
- [28] Y. Y. Lee, H. Ke, T. Y. Yen, H. H. Huang, and H. H. Chen, “Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement,” *J Assoc Inf Sci Technol*, vol. 71, no. 6, pp. 657–670, Jun. 2020, doi: 10.1002/ASI.24289.
- [29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, Jan. 2013, doi: 10.48550/arxiv.1301.3781.
- [30] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation,” pp. 1532–1543, Accessed: Jul. 22, 2022. [Online]. Available: <http://nlp>.
- [31] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of Tricks for Efficient Text Classification.” [Online]. Available: <https://github.com/facebookresearch/>
- [32] T. Mikolov, E. Grave, P. Bojanowski, C. Puhersch, and A. Joulin, “Advances in Pre-Training Distributed Word Representations,” *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pp. 52–55, Dec. 2017, doi: 10.48550/arxiv.1712.09405.
- [33] R. Speer, J. Chin, and C. Havasi, “ConceptNet 5.5: An Open Multilingual Graph of General Knowledge,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Dec. 2016, doi: 10.48550/arxiv.1612.03975.

- [34] M. Sanderson and J. Zobel, “Information retrieval system evaluation: Effort, sensitivity, and reliability,” *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 162–169, 2005, doi: 10.1145/1076034.1076064.
- [35] T. Mikolov, W.-T. Yih, and G. Zweig, “Linguistic Regularities in Continuous Space Word Representations”, Accessed: Jul. 31, 2022. [Online]. Available: <http://research.microsoft.com/en->
- [36] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” *Adv Neural Inf Process Syst*, Oct. 2013, doi: 10.48550/arxiv.1310.4546.

Appendices

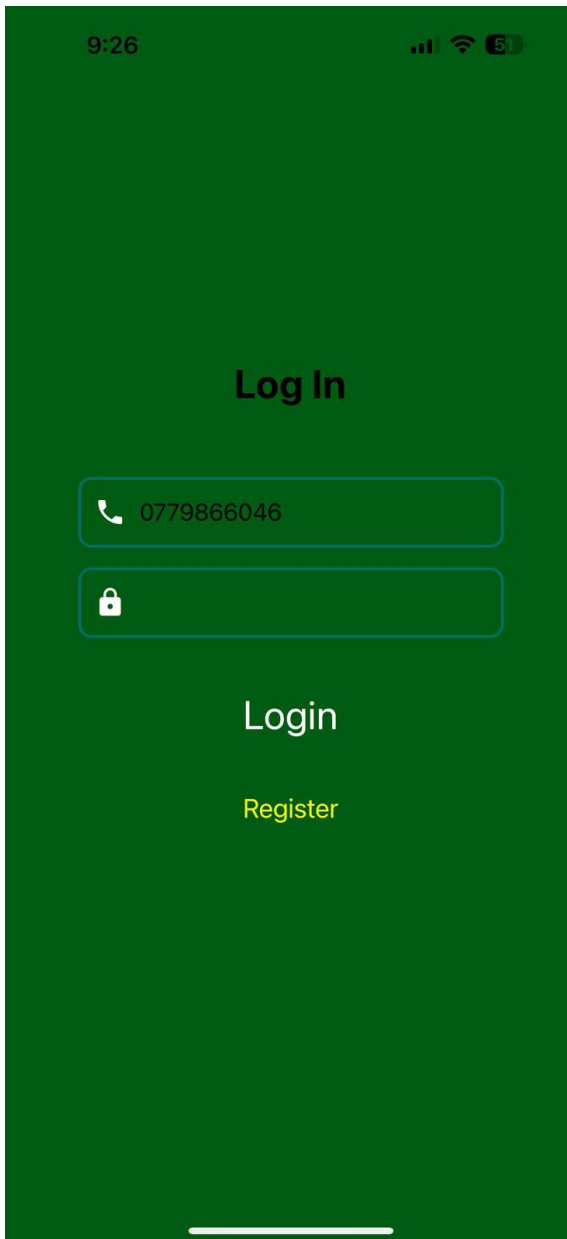


Figure 7: Mobile app Login



Figure 8: Mobile app dashboard

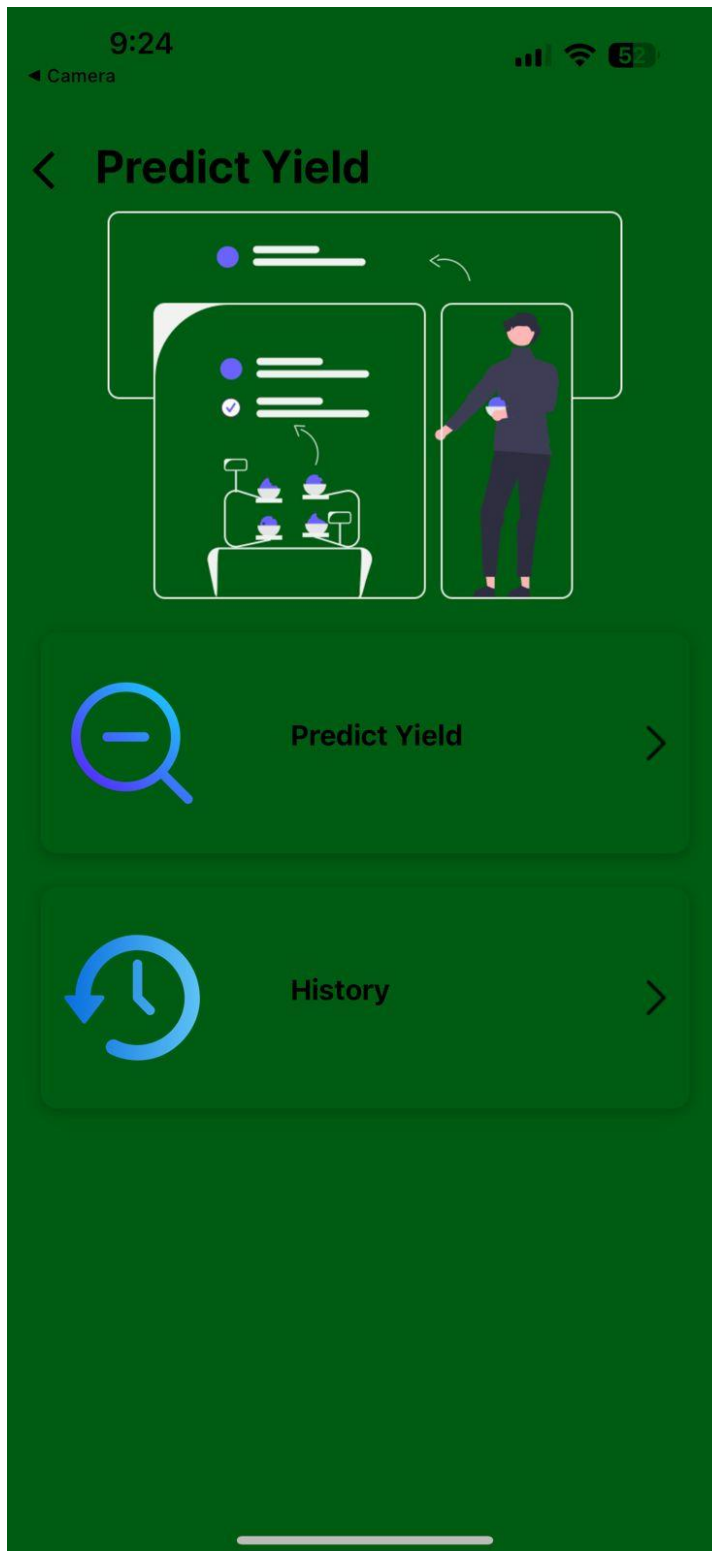


Figure 9: Mobile app yield prediction

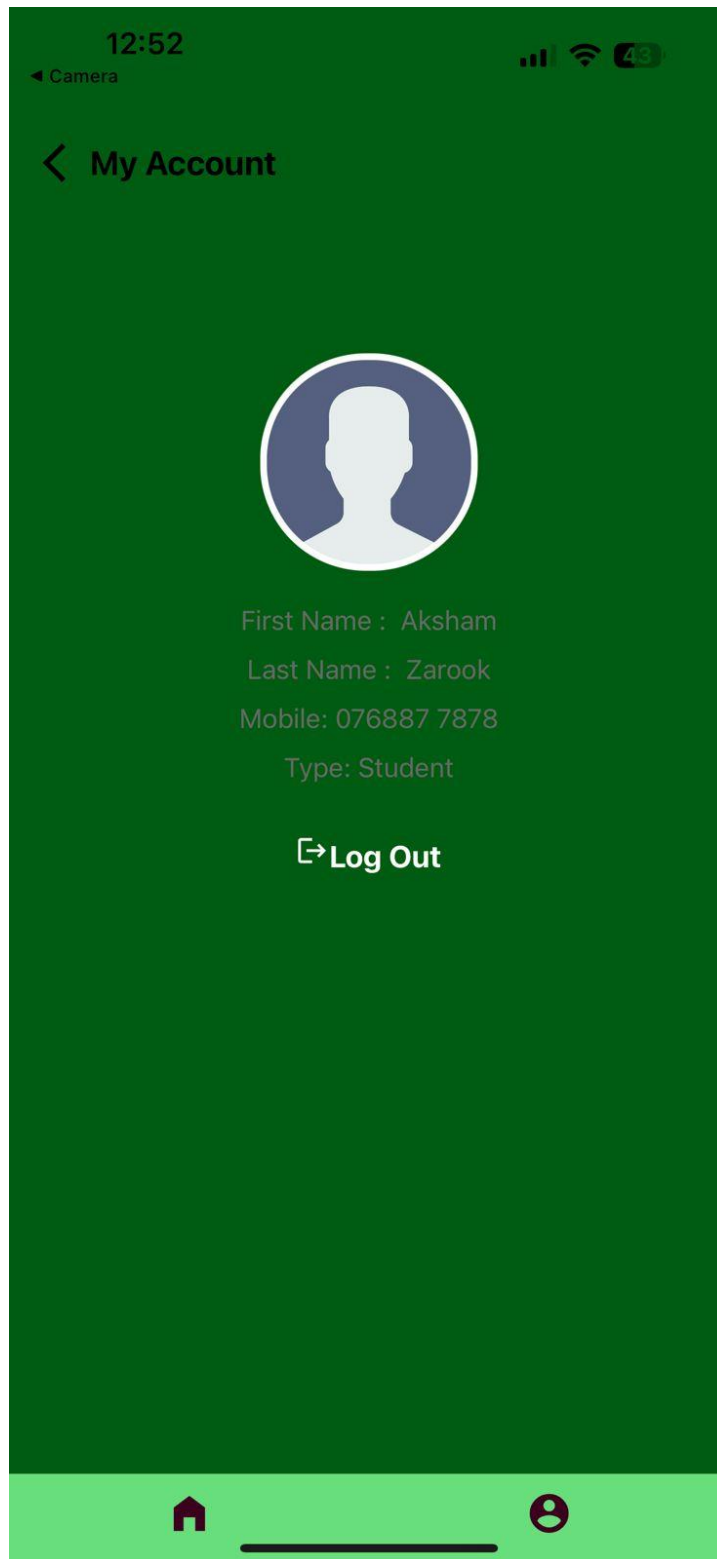


Figure 10: Mobile app Log out