



Project Report
on

Prediction of Suitable Crop Using Machine Learning

Submitted by

Project Members

Shranay Shahane 1032190916
Sakshi Dongre 1032191997
Nihaal Shetty 1032190927
Shivansh Kushwaha 1032180745

Under the Internal Guidance of

Dr. Himangi Pande

**School of Computer Engineering and Technology
MIT World Peace University, Kothrud,
Pune 411 038, Maharashtra - India
2022-2023**



Dr. Vishwanath Karad
MIT WORLD PEACE
UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY

C E R T I F I C A T E

This is to certify that,

Shranay Shahane
Sakshi Dongre
Nihaal Shetty
Shivansh Kushwaha

of BTech.(Computer Science & Engineering) have completed their project titled
“**Prediction of Suitable Crop Using Machine Learning System**” and have submitted this
Capstone Project Report towards fulfillment of the requirement for the
Degree-Bachelor of Computer Science & Engineering (BTech-CSE) for the academic
year 2022-2023.

[Dr. Himangi Pande]

Project Guide
School of CET
MIT World Peace University, Pune

[Dr. Vrushali Kulkarni]

Program Head
School of CET
MIT World Peace University, Pune

Internal

Examiner:

External

Examiner:

Date:

Acknowledgement

Any achievement, be it scholastic or otherwise does not depend solely on the individual efforts but on the guidance, encouragement and cooperation of intellectuals, elders and friends. A number of personalities, in their own capacities have helped me in carrying out this project. We would like to take an opportunity to thank them all.

First and foremost we express our sincere gratitude to the School of Computer Science & Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune for providing us the opportunity and support to present our project “Prediction of Suitable Crop Using Machine Learning System”.

We are very grateful to our project guide, Dr. Himangi Pande for her able guidance, regular source of encouragement and assistance throughout our project period. Her guidance helped us to bring the project to fruition and complete it in time.

Last, but not the least, we would like to thank our peers and friends who provided me with valuable suggestions to improve my project.

Abstract

Agriculture is the main source of income for most developing countries. Modern agriculture is an ever-expanding approach to agricultural progress and agricultural technology. Meeting the evolving demands of our planet and the expectations of dealers, customers and others is becoming a challenge for farmers. The impact of climate change in India, most of the crops have been severely impacted in terms of their performance over the last 20 years. Predicting crop yields before harvest will help decision-makers and farmers take appropriate measures for marketing and storage.

This project will help farmers know the yield of their crop before farming in the agricultural field and thereby help them make the right decision. It tries to solve the problem by building a prototype of the interactive prediction system. The implementation of such a system with easy-to-use web-based application and learning algorithms machine will be performed. Predicted results will be provided to farmers. So for this kind of data analysis in crop forecasting, there are different techniques or algorithms and with the help of these algorithms we can predict the crop yield. Analyzing all these problems and problems like temperature, humidity, ph, rainfall, etc., there is no right solution and technology to fix the situation we are in. In India, there are many ways to accelerate economic growth in agriculture.

Data mining is also useful for predicting crop yields. In general, data mining is the process of analyzing data from different perspectives and synthesizing that data into meaningful information. This project aims to find optimal crop prediction models to help farmers decide which crop types to grow based on climatic conditions and nutrients present in the soil. This paper compares popular algorithms such as K-Nearest Neighbor (KNN), Decision Tree, Random Forest Classifier, Naive Bayes, Support Vector Machine, XGBoost and Logistic Regression. Results reveal that Random Forest gives the highest accuracy among all.

Keywords— Agriculture, Machine Learning, Crop-prediction, Supervised Algorithms, Crop yield, Data Mining.

List of Figures

1. System Architecture.....	13
2. Machine Learning Module.....	14
3. Front-End Module.....	15
4. Activity Diagram.....	15
5. Steps involved in Methodology.....	17
6. Dataset Sample.....	17
7. Decision Tree.....	19
8. Support Vector Machine.....	20
9. Logistic Regression.....	20
10. Random Forest.....	21
11. K-Nearest Neighbour.....	22
12. Ensemble Modelling.....	22
13. Comparison of accuracies of different models.....	23
14. Accuracy of Tested models.....	23
15. Crop Feature Distribution.....	24
16. Web-based Application (i).....	25
17. Web-based Application (ii).....	25

List of Tables

1. Project Plan	16
-----------------------	----

Contents

Abstract	I
List of Figures	II
List of Tables	III

1	Introduction		1
	1.1	Project Statement	1
	1.2	Purpose	2
	1.3	Objectives/ Aim	2
2	Literature Survey		3
	2.1	Case 1	3
		2.1.1 Abstract	3
		2.1.2 Takeaways	3
		2.1.3 Research Gaps	3
	2.2	Case 2	3
		2.2.1 Abstract	3
		2.2.2 Takeaways	4
		2.2.3 Research Gaps	4
	2.3	Case 3	4
		2.3.1 Abstract	4
		2.3.2 Takeaways	4
		2.3.3 Research Gaps	4
	2.4	Case 4	5
		2.4.1 Abstract	5
		2.4.2 Takeaways	5
		2.4.3 Research Gaps	5
	2.5	Case 5	5
		2.5.1 Abstract	5

		2.5.2	Prospective Architecture	6
		2.5.3	Takeaways	6
		2.5.4	Research Gaps	6
	2.6	Case 6		6
		2.6.1	Abstract	6
		2.6.2	Takeaways	7
		2.6.3	Research Gaps	7
	2.7	Case 7		7
		2.7.1	Abstract	7
		2.7.2	Takeaways	7
		2.7.3	Research Gaps	8
	2.8	Case 8		8
		2.8.1	Abstract	8
		2.8.2	Takeaways	8
		2.8.3	Research Gaps	8
	2.9	Case 9		8
		2.9.1	Abstract	9
		2.9.2	Takeaways	9
		2.9.3	Research Gaps	9
	2.10	Case 10		9
		2.10.1	Abstract	9
		2.10.2	Takeaways	9
		2.10.3	Research Gaps	10
	2.11	Case 11		10
		2.11.1	Abstract	10
		2.11.2	Takeaways	10
		2.11.3	Research Gaps	10

	2.1 2	Case 12	10
		2.12.1 Abstract	10
		2.12.2 Takeaways	11
		2.12.3 Research Gaps	11
3		Problem Statement	11
	3.1	Project Scope	11
	3.2	Project Assumptions	11
4		Project Requirements	12
	4.1	Hardware and Software Requirements	12
	4.2	Risk Management	12
	4.3	Functional Specifications	12
5		System Analysis Proposed Architecture	13
	5.1	Overview of System Design	13
	5.2	System Architecture	13
	5.3	Modules of the Project	14
	5.4	UML Diagram	15
6		Project Plan	16
7		Implementation	16
	7.1	Methodology	17
		7.1.1 Collecting the Raw Data	17
		7.1.2 Data Preprocessing	17
		7.1.3 Train and Test Split	18
		7.1.4 Fitting the Model	18
		7.1.5 Checking the score over a training set	18
		7.1.6 Predicting a model	18
		7.1.7 Accuracy	18
	7.2	Algorithm	18
		7.2.1 Decision Tree	18

		7.2.2	Naive Bayes	19
		7.2.3	Support Vector Machine	19
		7.2.4	Logistic Regression	20
		7.2.5	Random Forest	20
		7.2.6	XGBoost	21
		7.2.7	K- Nearest Neighbour (KNN)	21
		7.2.8	Ensemble Modelling	22
	7.3	Other Implementations		22
		7.3.1	Dash	22
8	Performance Evaluation			23
	8.1	Accuracy Comparison of all different models		23
9	Result and Analysis			24
	9.1	Crop Feature Distribution		24
	9.2	Web Application (Screenshots)		24
	Conclusion			25
	Future prospects of the project			26
	References			26
	A. Base Paper(s)			27
	B. Plagiarism Report from any open source			28

Chapter 1

Introduction

Since its start, agriculture has been the main activity in every society and civilization that has existed throughout human history. It is not only a huge part of the expanding economy, but it is also crucial to our survival. It is the backbone of the Indian economy. In India, agricultural yields are mainly dependent on weather conditions. Cultivation of rice depends mainly on rainfall. Timely advice to predict future crop productivity and analysis is implemented to help farmers maximize crop yields.

In the past, Farmer predicted yields based on their experience with the previous year's yields. Therefore, there are various techniques or algorithms for this kind of data analysis in crop forecasting, with the help of those algorithms the crop yield can be predicted. Machine learning is a valuable decision-making tool for predicting agricultural yields and deciding what crops to sow and what to do during the growing season. In recent years, agriculture has used machine learning techniques. Crop forecasting is one of the complex challenges in agriculture and to date, several models have been developed and tested. Since agricultural production is affected by many factors such as temperature, humidity, ph, rainfall, etc. this challenge requires the use of multiple datasets. This implies that forecasting agricultural productivity is not a simple process; rather, it involves a series of complex procedures. Crop yield prediction methods can now reasonably estimate actual yield, although more excellent yield prediction performance is still desired.

The project aims to compare various algorithms like KNN, Decision Tree, Random Forest, XGBoost, Support Vector Machine, Logistic Regression and Naive Bayes on the dataset containing 22 varieties of crops. The results reveal that the suggested machine learning technique's effectiveness is compared to the best accuracy.

1.1 Project Statement

This project is based on curbing the problem faced by the farmers as well as providing the accurate level of the harvest they can expect from the crop they have grown depending on the dependent factors like temperature, rainfall, etc. This project is mainly developed to help farmers so that this may help them in analysis of the harvest of the crop. Farmers are facing losses in the crop yield due improper knowledge of the crop and the natural factors that are affecting them. In this project we analyze the factors and predict a graph that shows the crop's yield well before the harvest.

1.2 Purpose

It has been observed that farmers face problems at the time of crop yield due to rapidly changing weather that affects crop yield. Reduced quality of harvest and hence less income for farmers. This project aims to improve the quality of the harvest, helping farmers earn more money. In this project, we have collected a data set of some crop-dependent factors. Using this data, predictions are obtained to show that the harvest of the crop is growing.

1.3 Objectives

The main objectives are:-

- a) To identify the features affecting the crop yields from the selected database.
- b) To test different machine learning algorithms on the selected database and perform a comparative analysis.
- c) To create a prediction model that will predict the most suitable crop for the user's given environmental conditions and soil type.
- d) To create a user-friendly web based application that will act as an interface between the user and the prediction engine.

Chapter 2

Literature Survey

Considering the importance of crop prediction, many proposals have been made in the past with the aim of improving the accuracy of crop prediction.

2.1 Case 1 : A Comprehensive Review on the Crop Prediction Algorithms by C.G. Anupama, C. Lakshmi (Science Direct, 2021)

2.1.1 Abstract

This paper discusses different papers with similar problem statements and different approaches they tried to solve the problem. Various Machine Learning algorithms including decision trees, random forest, knn, k means, ANN, etc have been discussed. The paper discusses the performance of different algorithms on different datasets in an attempt to make novel observations. Most of the datasets used were from government websites.

2.1.2 Takeaways

Clustering and Classification approaches are widely used for crop yield prediction models. Naive Bayes, Random Forest and Decision Trees are the most used machine learning approaches used in these projects. Most considered features in the datasets are rainfall, location, soil type, crops and area of land. Noisy Data can drastically affect the prediction results.

2.1.3 Research Gaps

The comparative analysis between different approaches and algorithms has not been performed. Most papers do not discuss or consider some important factors like Humidity, NPK values in soil, ph value of soil, etc. which can have a major impact on the prediction outcome. Most data was location dependent limiting the prediction to a specific district or state.

2.2 Case 2 : Crop Yield Prediction using Machine Learning Algorithm by Anakha Venugopal, Aparna S, Jinsu Mani, Rima Mathew, Prof. Vinu Williams (IJERT, 2021)

2.2.1 Abstract

In this paper, different machine learning approaches are used and a comparative analysis is performed on random forest, naive bayes and logistic regression. This research work was completed by creating an android application as the front end of the Machine Learning Model. The application makes an attempt to predict the estimated revenue generated by selling all the crop yield at the current market price for a single hectare of farmland.

2.2.2 Takeaways

Based on the results, it can be determined that the random forest algorithm was the best fit for the dataset with highest accuracy of 92.8%, followed by Naive Bayes with an accuracy of 91.5% and at last logistic regression with an accuracy of 87.8%. Factors like wind speed, rainfall, humidity, temperature were considered for the training of the Machine Learning Model. Weather API was used to get the real time weather conditions of the provided locations.

2.2.3 Research Gaps

As discussed in the future scope of the research paper, the soil data of farmland has not been yet introduced as one of the features which will have significant impact on the final results of the prediction engine. A few other algorithms like Decision Trees and Support Vector Machine which could potentially have provided better results were not tested. The data was location dependent limiting the model predictions to the state of Kerala.

2.3 Case 3 : Analysis of agricultural crop yield prediction using statistical techniques of machine learning by Janmejy Pant, R.P. Pant, Manoj Kumar Singh, Devesh Pratap Singh, Himanshu Pant (Science Direct, 2021)

2.3.1 Abstract

In this research work different machine algorithms are used to predict crop yield in India. Datasets for making predictions for four primary crops which are potatoes, rice, wheat and maize have been used. Machine learning Algorithms like Gradient Boosting Regressor, Random Forest Regressor, SVM and Decision Tree Regressor are used and experimented upon. Comparative Analysis between those different approaches have been performed.

2.3.2 Takeaways

The dataset includes features like rainfall, pesticide used, yield production of specific year, average temperature, etc. From the comparative analysis performed, it is pretty clear that the Decision Tree Regressor was the best approach for the given dataset with an accuracy of 96%. Based on the results, it can be said that potato is the most suitable crop in India. The test-train data was created by originally splitting the data in 70-30 ratio.

2.3.3 Research Gaps

The research work does not consider major factors like soil, humidity, ph values, etc. The research work also does not factor the land under cultivation in an attempt to generalise the result to the whole country which might give wrong results since the climate change could be vast over the country affecting crop production.

2.4 Case 4 : Prediction of the production of crops with respect to rainfall by Benny Antony (Science Direct, 2021)

2.4.1 Abstract

In this research work, yields of different major crops from different states are predicted using various algorithms and then the comparative analysis is done. The techniques used in this research are linear regression, polynomial regression, support vector machine, random forest and XGboost algorithm. MAE (Mean Absolute Error) is calculated for each method and crop and the approach with least MAE is considered the best approach for the given dataset.

2.4.2 Takeaways

According to the comparative analysis done in the research paper, decision tree, random forest and XGboost appear to be the best fitting machine learning algorithms for all the datasets each having slight differences in their MAE. This research work only focuses on measuring the impact of rainfall on various crops in various states. Support Vector Machine seems to have the highest MAE in all the individual cases.

2.4.3 Research Gaps

The research work only considers the impact of rain on the crops and not the other factors like the soil, temperature, humidity, etc. Production/Area is measured for individual crops and states which does not render the relationship between the climate or state and the crops. Hence, it cannot be determined which crops are more or less suitable for which environment. Rainfall is measured yearwise which might give wrong relational data between rainfall and rabi crops (Crops which do not require much water and are sown in winter and harvested in summer.)

2.5 Case 5 : A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction by Mamunur Rashid, Bifta Sama Bari, Yusri Yusup, Mohamad Anuar Kamaruddin, Nuzhat Khan (IEEE Access (Volume: 9))

2.5.1 Abstract

This article provides an exhaustive review on the use of machine learning algorithms to predict crop yield with special emphasis on palm oil yield prediction. Initially, the current status of palm oil yield around the world is presented, along with a brief discussion on the overview of widely used features and prediction algorithms. Then, the critical evaluation of the state-of-the-art machine learning-based crop yield prediction, machine learning application in the palm oil industry and comparative analysis of related studies are presented. Consequently, a detailed study of the advantages and difficulties related to machine learning-based crop yield prediction and proper identification of current and future challenges to the agricultural industry is presented. The potential solutions are additionally prescribed in order to alleviate existing problems in crop yield prediction.

2.5.2 Prospective Architecture

In order to predict the crop yield, a wide range of data including leaf and fruit information, irrigation information, soil properties, climatic information, etc. is collected in the first step. After data collection, the data is pre-processed for further analysis. Once the data is pre-processed, the entire dataset is split into a training and a testing set. Different ML based regression and classification algorithms are then employed in the model's training phase. If the performance of the trained model is not satisfied, the parameter of the prediction model is optimized. After getting the threshold performance, the trained model is tested through the testing dataset. There are some crucial agrarian factors that have significant impact on crop yield prediction and those agrarian factors include disease recognition and management, mapping and plant counting, plant growth and nutrition level monitoring and natural disaster. Finally, the output from ML is adjusted with the agrarian factors' output to get the precise palm oil yield prediction.

2.5.3 Takeaways

They have done a lot of estimations to find which crop should they do prediction for, then which feature affects the most for that crop yield, and then which machine learning algorithm is better to used to do the crop yield predictions. The most promising conventional ML architectures are LR, RF and NN. Besides these algorithms, some DL models, including DNN, CNN and LSTM, are also employed in crop yield estimation.

2.5.4 Research Gap

As they have used a lot of references, this paper is a little complex to understand because they have compared their problem statement with a lot of other things. No machine learning algorithms are applied here for crop yield prediction; they only talked about which ML algorithm is better to use for prediction. No information about the dataset used is given. The crop yield prediction is based on satellite imaging rather than datasets of various factors affecting the crop yield. The paper does not include any comparative analysis.

2.6 Case 6 : An Approach for Prediction of Crop Yield using Machine Learning and Big Data Techniques by Kodimalar Palanivel, Chellammal Surianarayanan (IAEME Publication , 2019)

2.6.1 Abstract

The productivity of the crops is significantly influenced by weather conditions. Early prediction of yield would facilitate the farmers to make precautionary actions to improve productivity. Early prediction is possible through collection of previous experience of the farmers, weather conditions and other influencing factors and; store it in a large database. In this paper, an investigation has been performed on how various machine learning algorithms are useful in prediction of crop yield. An approach has been proposed for prediction of crop yield using machine learning techniques in the big data computing paradigm. They have used Linear Regression, Artificial Neural Networks and Support Vector Machine for their

prediction and the metrics they used are Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

2.6.2 Takeaways

An investigation has been performed on how various machine learning algorithms are useful in prediction of crop yield. The expert system such as fuzzy logic is based on logical rules to predict the yield. The performance metrics of the machine learning algorithms such as root mean square error are studied. Along with machine learning algorithms for prediction, it is planned to study the impact of big data techniques in the prediction of crop yield.

2.6.3 Research Gaps

Model was very complex, that's why diverse machine learning techniques were adopted. This paper only showcases the different techniques used for prediction but fails to perform a comparative analysis. Limited ML algorithms used. Better algorithms could have been more advantageous.

2.7 Case 7 : Machine learning methods for crop yield prediction and climate change impact assessment in agriculture by Andrew Crane-Droesch^{2,1} (IOP Publishing Ltd.)

2.7.1 Abstract

Crop yields are critically dependent on weather. A growing empirical literature models this relationship in order to project climate change impacts on the sector. This paper describes an approach to yield modeling that uses a semiparametric variant of a deep neural network, which can simultaneously account for complex nonlinear relationships in high-dimensional datasets, as well as known parametric structure and unobserved cross-sectional heterogeneity. Using data on corn yield from the US Midwest, we show that this approach outperforms both classical statistical methods and fully-nonparametric neural networks in predicting yields of years withheld during model training. Using scenarios from a suite of climate models, we show large negative impacts of climate change on corn yield, but less severe than impacts projected using classical statistical methods.

2.7.2 Takeaways

This paper focuses on the latter—yield prediction from weather. In this paper, two different algorithms are used and that is Semiparametric Neural Networks (SNN) and Ordinary Least Square (OLS). The error rate was measured with the help of Mean Square Error (MSE) and Coefficient of Determination (R^2). This also demonstrates comparative study of algorithms with the amount of accuracy and r^2 score they obtained and suggesting the best algorithm which could be implemented for predicting the crop and yield respectively. This paper is focused on supervised ML—used for prediction—rather than unsupervised ML, which is used to discover structure in unlabeled data.)

2.7.3 Research Gaps

This paper only took into consideration weather. Parameters like Rainfall, Temperature, Pesticides, etc. were not used. In this paper the dataset was made from a particular selected region that is the US Midwest. Limited algorithms used. Better algorithms could have been more advantageous.

2.8 Case 8 : Machine Learning Approaches to Corn Yield Estimation Using Satellite Images and Climate Data: A Case of Iowa State by Kim Nari, Lee Yang-Won

2.8.1 Abstract

Remote sensing data has been widely used in the estimation of crop yields by employing statistical methods such as regression models. Machine learning, which is an efficient empirical method for classification and prediction, is another approach to crop yield estimation. This paper described the corn yield estimation in Iowa State using four machine learning approaches such as SVM (Support Vector Machine), RF (Random Forest), ERT (Extremely Randomized Trees) and DL (Deep Learning). Also, comparisons of the validation statistics among them were presented. To examine the seasonal sensitivities of the corn yields, three period groups were set up: (1) MJJAS (May to September), (2) JA (July and August) and (3) OC (optimal combination of month). In this paper, they used satellite images from MODIS and the climate reanalysis data created by PRISM (Parameter-Elevation Regressions on Independent Slopes Model) for the machine learning analyses.

2.8.2 Takeaways

This paper described the estimation of corn yields in Iowa State using four machine learning techniques such as SVM, RF, ERT and DL, and presented the comparisons of the validation statistics among them. In overall, the DL method showed the highest accuracies in terms of the correlation coefficient for all the period groups. Remote sensing data has been widely used in the estimation of crop yields by employing statistical methods. Satellite images from MODIS are used and the climate reanalysis data is created by PRISM for the machine learning analyses.

2.8.3 Research Gaps

The model is very complex with many parameters and shows poor predictive performance by overreacting to minor fluctuations in the dataset; which resulted in an overfitting problem. The crop yield prediction is based on satellite imaging rather than datasets of various factors affecting the crop yield. In this paper the dataset was made from a particular selected region that is the US Midwest.

2.9 Case 9 : Performance Evaluation of Machine Learning Techniques for Mustard Crop Yield Prediction from Soil Analysis by Vaishali Pandith , Haneet Kour , Surjeet Singh , Jatinder Manhas , and Vinod Sharma.

2.9.1 Abstract

A key factor influencing crop production prediction is soil. By creating advanced plans, farmers and soil analyzers can use analysis of soil nutrients to increase crop productivity. In this study, a variety of machine learning algorithms have been applied to forecast mustard crop output from soil data in advance. The Department of Agriculture, Talab Tillo, Jammu, provided the information for the experimental setup, which included soil samples from various districts in the Jammu region for the mustard crop. Five supervised machine learning methods have been used in the current study: K-Nearest Neighbor (KNN), Naive Bayes, Multinomial Logistic Regression, Artificial Neural Network (ANN), and Random Forest. Five factors, including accuracy, recall, precision, specificity and f-score are used to evaluate each technique being studied.

2.9.2 Takeaways

Naive Bayes predicted the lowest accuracy, 72.33%, while KNN and random forest projected the best accuracy, 88.67% and 94.13%, respectively. In terms of precision, Logistic Regression predicted the lowest value, which was 24.17%, while ANN predicted the greatest value, which was 99.94%. Except for Naive Bayes, all of the classifiers under investigation predicted recall values of greater than 90%. This indicates that Logistic regression had a high false positive rate with a low true negative rate, whereas Naive Bayes had the largest false negative rate. Higher specificities of 99.78% and 80.72% were produced by ANN and KNN, respectively, as well as the highest f-scores of 0.9976 and 0.8405.

2.9.3 Research Gap

In the future, a big data environment could be used to estimate crop yields using extensive soil data. In the event that a poor crop yield is predicted, fertilizer suggestions can also be put into practise based on the yield prediction results to assist soil analyzers and farmers in making the appropriate decisions.

2.10 Case 10 Crop Yield Prediction Using Deep Neural Networks

2.10.1 Abstract

Deep Neural Network is used for prediction.

2.10.2 Takeaway

In the 2018 Syngenta Crop Challenge, we presented a machine learning approach for crop yield prediction that outperformed the competition utilising massive datasets of corn hybrids. Based on genotype and environmental data, the method used deep neural networks to predict yield (including yield, check yield, and yield difference). From historical data, the carefully crafted deep neural networks were able to infer nonlinear and complex relationships between genes, environmental factors, and their interactions and predict yields for new hybrids planted in unfamiliar locations with known weather conditions fairly accurately.

2.10.3 Research Gap

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 GE interactions, it is challenging to develop testable hypotheses that might offer biological insights because of the model's complicated model structure. We used the backpropagation method to do feature selection based on the trained DNN model to lessen the model's black box nature. The feature selection method was successful in identifying significant features and showed that environmental conditions affected crop production more so than genotype.

2.11 Case 11 An Ensemble Algorithm for Crop Yield Prediction by Mummaleti Keerthana, K J M Meghana, Siginamsetty Pravallika, Modepalli Kavitha.

2.11.1 Abstract

This study examines the application of ensemble approaches for predicting the crop type based on location-specific characteristics. Machine learning may predict the output from a set of input parameters using supervised or unsupervised learning methods. We must create an acceptable and satisfying function utilising a set of variables that will represent the output (the desired variable) using the provided input variables or parameters in order to obtain the required output parameter. The ensemble (combination) of two machine learning methods is included in this, which increases the accuracy of crop yield forecast.

2.11.2 Takeaway

Created a technique for predicting agricultural yield using previously gathered data. With the help of some machine learning algorithms, this has been resolved.

To forecast the outcome with a higher rate of accuracy, an ensemble of decision tree regression and an AdaBoost regression is utilised in this case. As opposed to other algorithms, decision trees will improve accuracy, as we have realised.

Decision trees by themselves do not produce many correct results, which leads to a subpar output. AdaBoost regressor is ensembled to strengthen the decision tree's weak learner.

2.11.3 Research Gap

We can collectively combine a larger number of algorithms to improve the current system. Additionally, we have access to sophisticated complex algorithms that have greater prediction capacity, improving accuracy. These techniques improve accuracy and produce results that farmers in the agriculture sector find appealing.

2.12 Case 12 ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUES by D Ramesh and Vishnu Vardhan

2.12.1 Abstract

Various data mining methods are employed and assessed in agriculture to forecast crop yield

for the following year. For the chosen area, the East Godavari district of Andhra Pradesh in India, this research gives a brief investigation of crop yield prediction using Multiple Linear Regression (MLR) technique and Density based clustering technique.

2.12.2 Takeaway

On existing data, the statistical model's Multiple Linear Regression approach is first used. Utilizing the density-based clustering technique from data mining, the outcomes thus acquired were confirmed and examined.

2.12.3 Research Gap

In this paper dataset was made from a particular selected region that is East Godavari District and hence the dataset was small.

Chapter 3

Problem Statement

India is an agricultural country. Yield of each crop depends on its dependent factors. It is very important to predict the yield of a crop to help farmers. Crop Yield Prediction is predicting the yield of a crop in future based on the dependent factors. Crop yield is dependent on factors like rainfall, pressure, temperature and area or the geographical location. This is achieved by :-

- a) Designing a system to predict crop yield.
- b) Providing graphical user interface to view prediction results.

3.1 Project Scope

In the project, we introduce a scalable, accurate and low-cost method for predicting crop ideal for the given environment using various factors and machine learning. Our machine learning approach can predict the crop yield with high spatial resolution. Machine learning algorithms like logistic regression algorithm, KNN, Naive Bayes, Random forest, Decision Tree, XGBoost and SVM algorithm are used to predict crop yield based on factors like temperature, rainfall, ph and humidity.

3.2 Project Assumptions

In this project, we assume that there will be no real time factor like unexpected rainfall, change in temperature levels, pest attacks, crop diseases or any other environmental factors that will affect the production of crops at any stage in the crop lifecycle after sowing the seeds. We have also not considered factors like fertilizers, pesticides, etc. that might affect the crop production.

Chapter 4

Project Requirements

4.1 Hardware and Software Requirements

Hardware Requirements :

Machine with Python Environment

Software Requirements :

Datasets of crops and ideal environment to grow these crops.

<https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset>

4.2 Risk Management

Risk is inherent in almost everything we do and is definitely associated with any software implementation project. A risk is something that “may happen”, implying a probability of less than 100%, and if it does transpire, will have an adverse impact on the project. If it has a probability of 100%, in other words, it occurs – then it becomes an issue. Such an issue is handled differently to a risk.

An effective methodology approach addresses risk management in four stages:

Stage 1: Identification

Stage 2: Quantification

Stage 3: Response

Stage 4: Control

4.3 Functional Specifications

The Conditions of Function The functions and actions that a system must be capable of doing are specified. Functional specifications should detail the tasks carried out by certain modules and the system's processes.

Chapter 5

System Analysis Proposed Architecture

5.1 Overview of System design

Design is a meaningful engineering representation of something that is to be built. It is the most crucial phase in the development of a system. Software design is a process through which the requirements are translated into a representation of software. Design is a place where design is fostered in software Engineering. Based on the user requirements and the detailed analysis of the existing system, the new system must be designed. This is the phase of system designing. Design is the perfect way to accurately translate a customer requirement in the finished software product. Design creates a representation or model, provides details about software data structure, architecture, interfaces and components that are necessary to implement a system. The logical system design arrived at as a result of systems analysis is converted into physical system design.

5.2 System Architecture

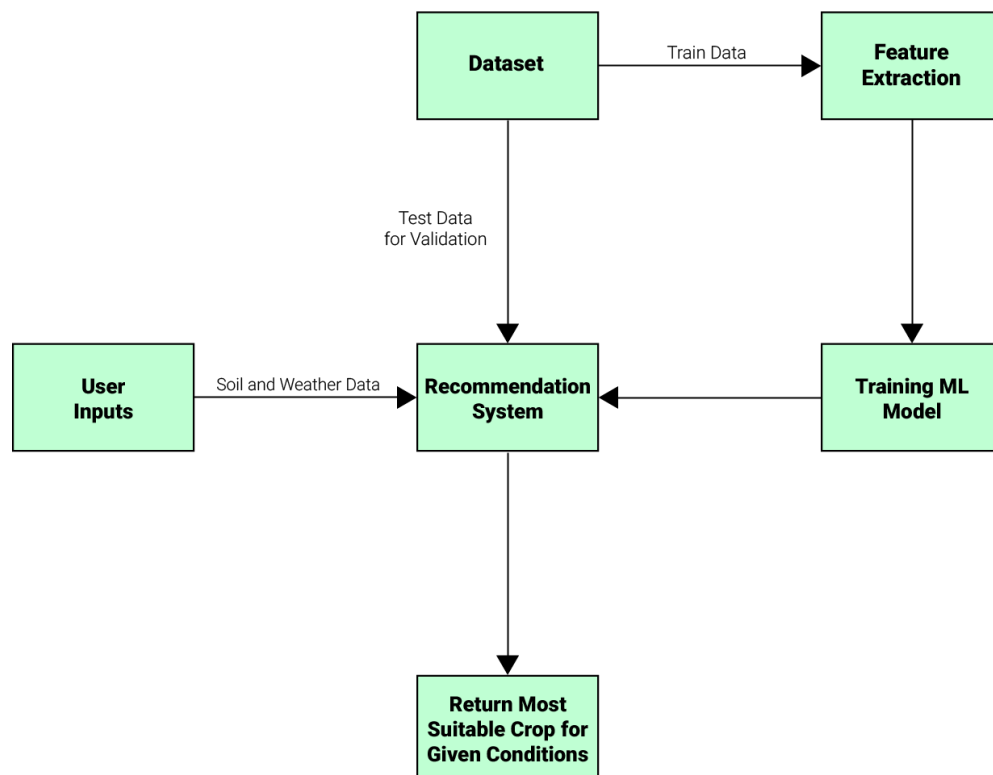


Fig 1 : System architecture

5.3 Modules of the Project

This Project Consists of 2 major modules:

1. Machine Learning Module: This module consists of the core decision system and trained machine learning model which will make the predictions. This module will include tasks like Data Preprocessing, Implementation of machine learning algorithms, testing trained models against test data sets, etc.

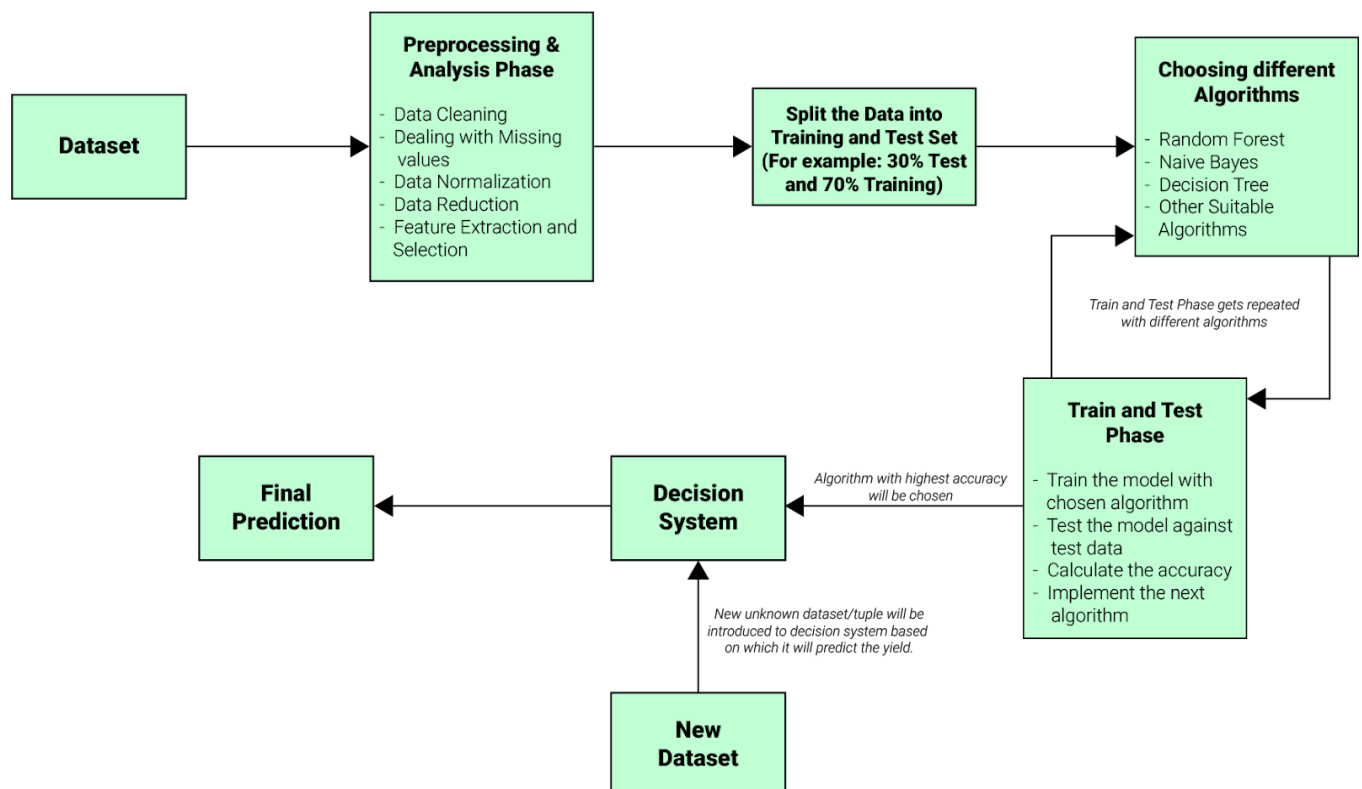


Fig 2 : Machine learning module

2. Front End Module: This module will include a web based application that will be user friendly and acts as an interface between the machine learning module and the user. This module will include tasks like creating a web page, connecting the web page to the machine learning module, etc.

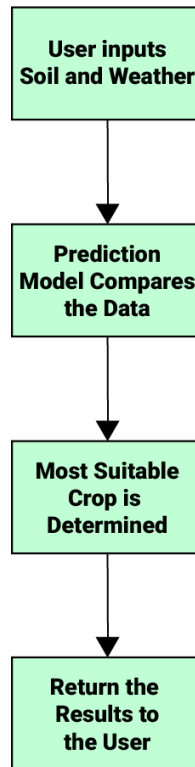


Fig 3 : Front-End Module

5.4 UML Diagram

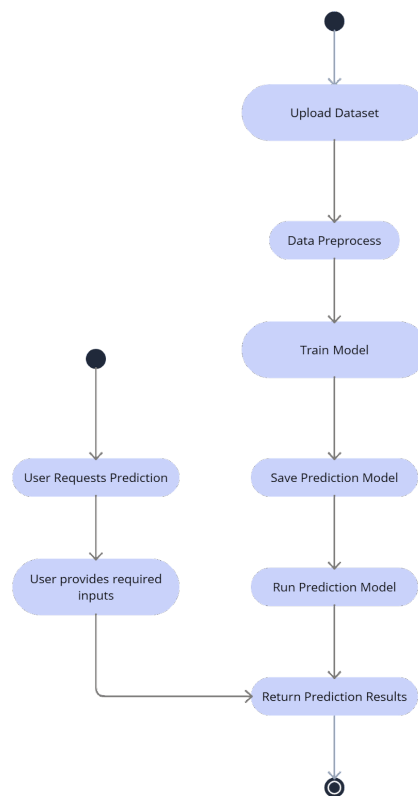


Fig 4 : Activity Diagram

Chapter 6

Project Plan

Sr. No.	Activity	Month				
		1	2	3	4	5
1	Feasibility Review					
2	Literature Survey					
3	Implementation					
4	Testing and Validation					
5	Documentation					

Table 1 : Project Plan

Chapter 7

Implementation

The implementation phase of the project is where the detailed design is actually transformed into working code. Aim of the phase is to translate the design into the best possible solution in a suitable programming language. This chapter covers the implementation aspects of the project, giving details of the programming language and development environment used. The implementation stage requires the following tasks:-

- Careful planning.
- Investigation of system and constraints.
- Design of methods to achieve the changeover.
- Evaluation of the changeover method.
- Correct decisions regarding selection of the platform.
- Appropriate selection of the language for application development.

7.1 Methodology

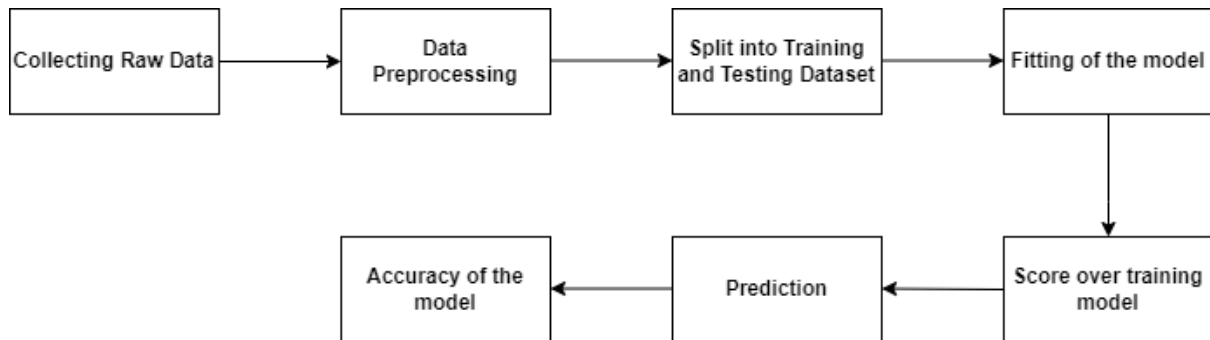


Fig 5 : Steps involved in Methodology

7.1.1 Collecting the Raw Data

Data collection means pooling data by scraping, capturing, and loading it from multiple sources, including offline and online sources. Since data collection is a way of tracking past events, data analysis can be used to find recurring patterns. The ‘Crop Recommendation’ dataset is collected from the Kaggle website. The dataset takes into account 22 different crops as class labels and 7 features- (i) Nitrogen content ratio (N) (ii) Phosphorus content ratio (P) (iii) Potassium content ratio (K) in the soil, (iv) Temperature expressed in degree Celsius (v) Percentage of Relative Humidity (vi) pH value and (vii) Rainfall measured in millimeters.

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

Fig 6 : Dataset Sample

7.1.2 Data Preprocessing

The process of modifying raw data in the form of learning algorithms to generate insights or predict outcomes is called data preprocessing. In this project, the data processing method is to find missing values. Getting every data point for every record in a data set is difficult. Blank

cells, values like zero, or certain characters like question marks can all indicate missing data. The dataset used in the project didn't have any missing values.

7.1.3 Train and Test Split

This is done by splitting the dataset into training and test datasets using the `train_test_split()` method of the scikit learn module. The 2200 data in the dataset were split into training dataset 1760 with 80% of the dataset and test dataset 440 with 20% of the dataset.

7.1.4 Fitting the model

Modifying model parameters to improve accuracy is called fitting. To build a machine learning model, you apply an algorithm to data with known target variables. Model accuracy is determined by comparing the model results to the actual observed values of the target variable. Model fitting is the ability of a machine learning model to generalize data comparable to that with which it was trained. For unknown inputs, a good model fit is one that correctly approximates the output.

7.1.5 Checking the score over a training dataset

Scoring, also often called prediction, is the creation of values from new input data using a trained machine learning model. Using the `model.score()` method, which computes the score for each model on the training set, shows how well the model has learned.

7.1.6 Predicting the model

When predicting the likelihood of a particular outcome, "prediction" refers to the outcome after the algorithm has been trained on previous datasets and applied to new data. Predicting the model using `predict()` method using test feature dataset. It has given the output as an array of predicted values.

7.1.7 Accuracy

The number of correct predictions divided by the total number of predictions is called the model's accuracy. The accuracy of the model is calculated using the `accuracy_score()` method of scikit learn metrics module.

$$\text{Accuracy} = [\text{TP} + \text{TN}] / [\text{TP} + \text{TN} + \text{FP} + \text{FN}]$$

where TP-True Positive; FP-False Positive; TN-True Negative; FN-False Negative

7.2 Algorithm

7.2.1 Decision Tree

Decision Tree is a supervised learning technique used for both classification and regression problems where each path is a set of decisions leading to a class. With high accuracy,

decision trees are capable of handling high-dimensional data. It's a flowchart diagram-style representation that closely parallels human-level thinking. As a result, decision trees are simple to explain and apprehend.

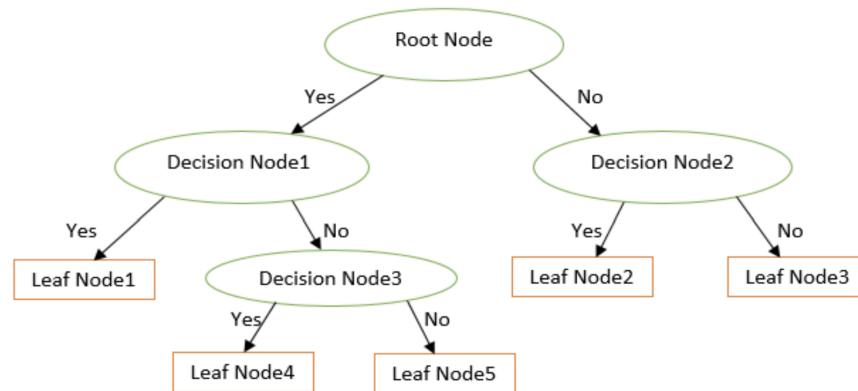


Fig 7 : Decision tree

7.2.2 Naive Bayes

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. The formula for Bayes' theorem is given as:

$$P(A|B) = \{P(B|A) \cdot P(A)\} / P(B)$$

7.2.3 Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

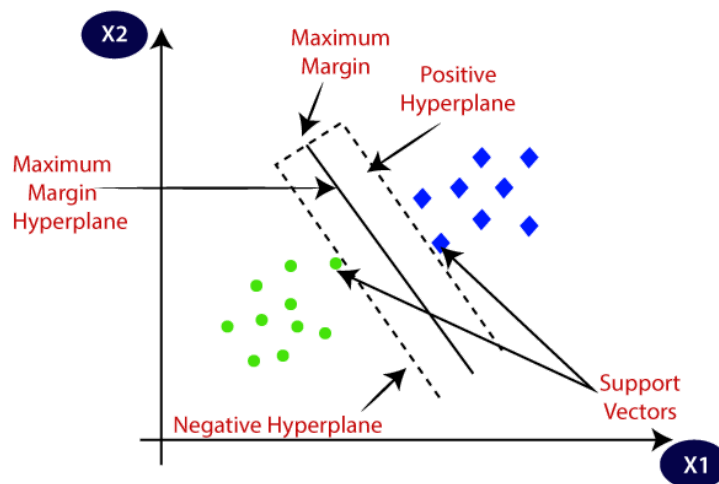


Fig 8 : Support Vector Machine

7.2.4 Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

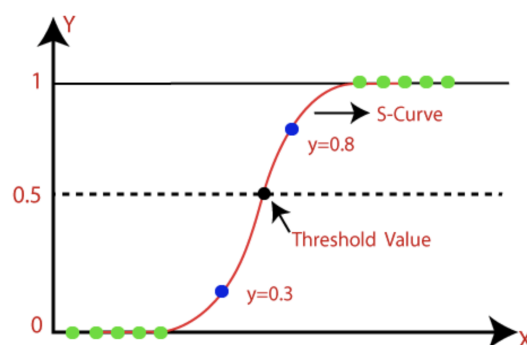


Fig 9 : Logistic regression

7.2.5 Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Instead

of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

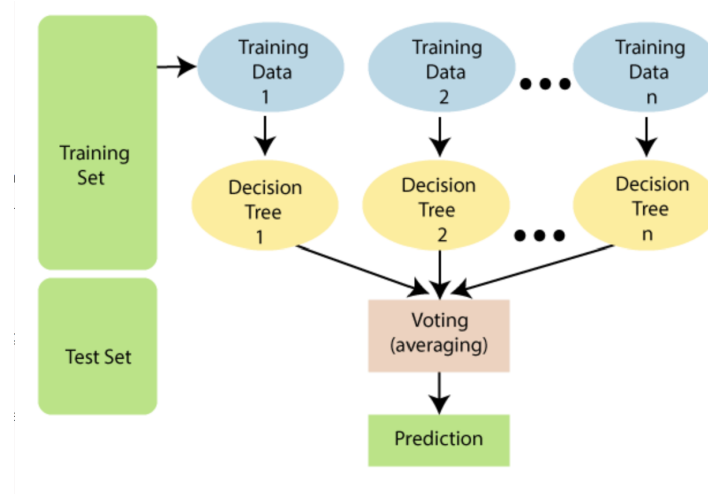


Fig 10 : Random Forest

7.2.6 XGBoost

Gradient boosted decision trees are implemented by the XGBoost library of Python, intended for speed and execution, which is the most important aspect of Machine Learning. This is an AI method utilized in classification and regression assignments, among others. It gives an expectation model as a troupe of feeble forecast models, commonly called decision trees.

7.2.7 KNN (K-Nearest Neighbour)

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. KNN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K- NN algorithm. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

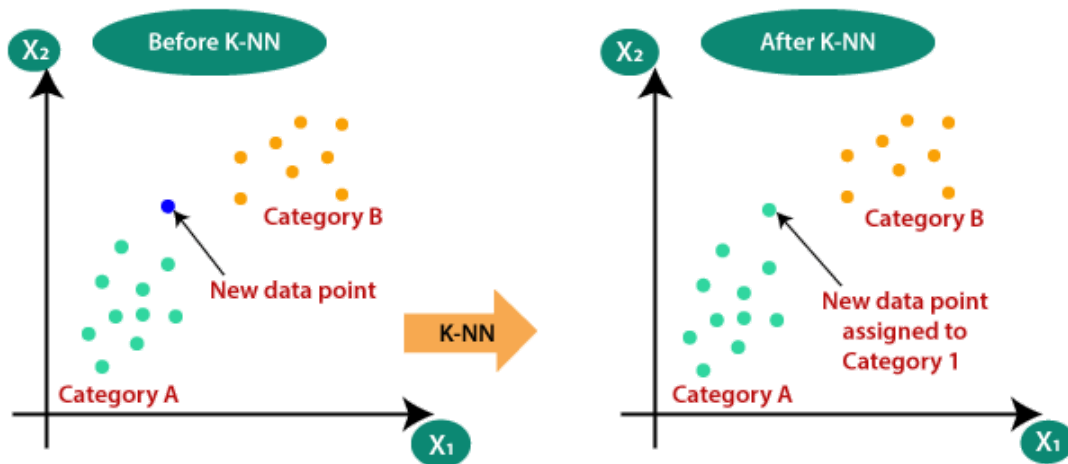


Fig 11 : K-NN

7.2.8 Ensemble Model

Ensemble modelling is the act of executing two or more related but separate analytical models and then combining the results into a single model. Ensemble Models are used in order to increase the precision of predictive analytics and data mining applications.

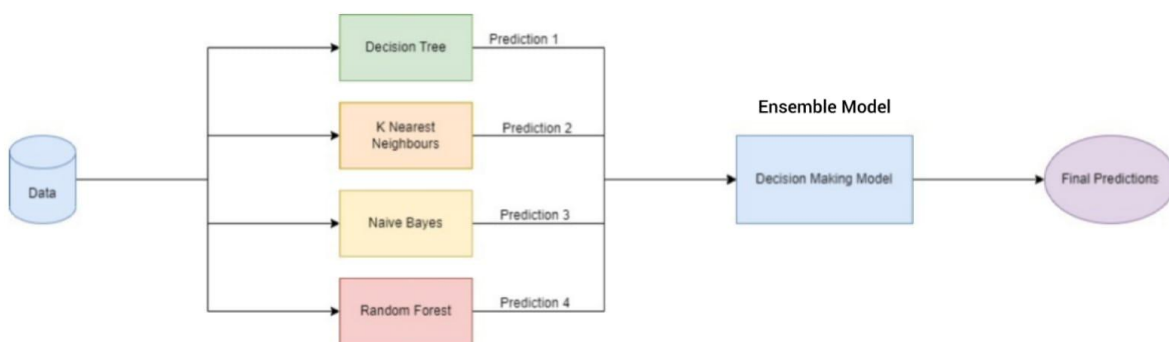


Fig 12 : Ensemble modelling

7.3 Other Implementation

7.3.1 Dash

Dash is an open-source Python framework used for building analytical web applications. It is a powerful library that simplifies the development of data-driven applications. It's especially useful for Python data scientists who aren't very familiar with web development. Users can create amazing dashboards in their browser using dash. Dash apps consist of a Flask server that communicates with front-end React components using JSON packets over HTTP

requests. Dash applications are written purely in python, so NO HTML or JavaScript is necessary.

Chapter 8

Performance Evaluation

8.1 Accuracy Comparison of all Models

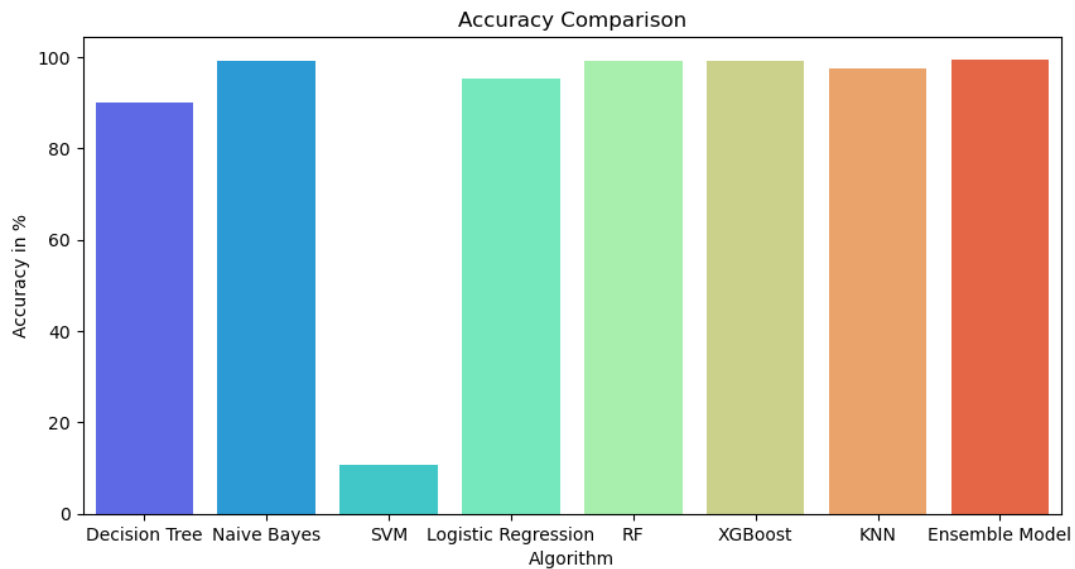


Fig 13 : Comparison of accuracies of different models

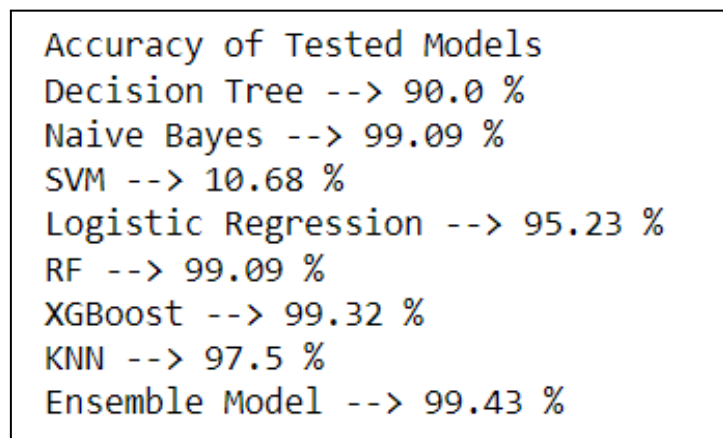


Fig 14 : Accuracy of Tested models

Chapter 9

Result and Analysis

9.1 Crop Feature Distribution

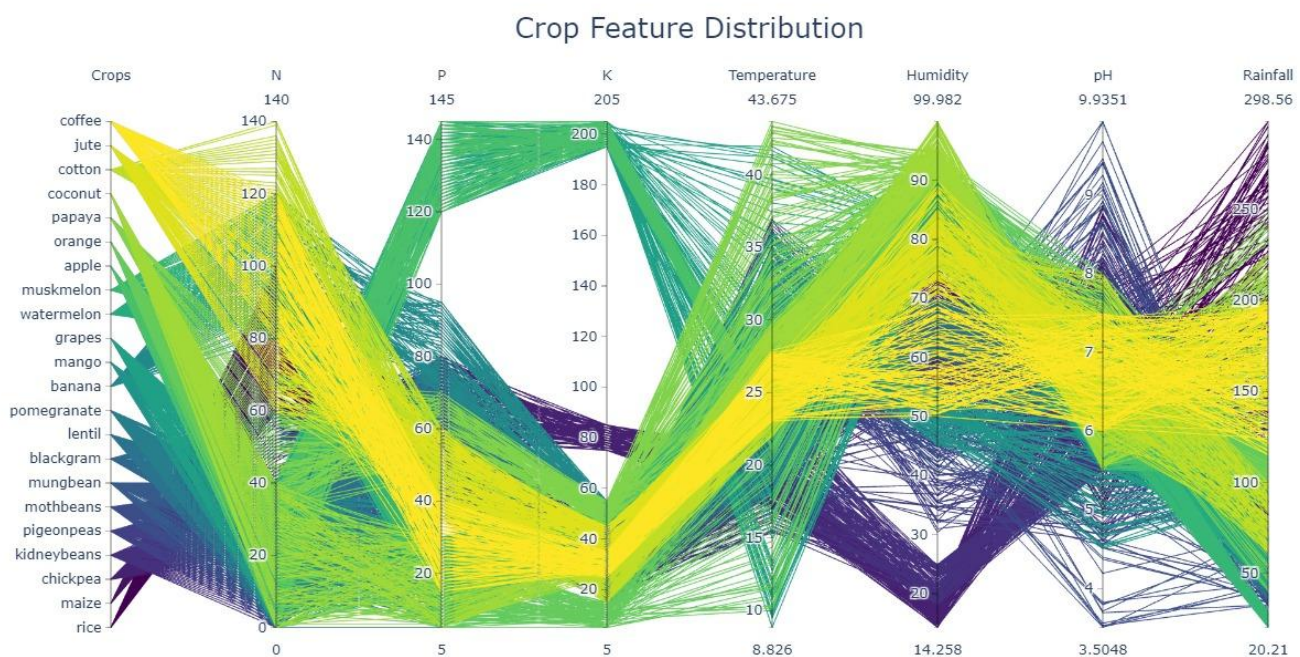


Fig 15 : Crop Feature Distribution

9.2 Web Application

Crop recommendation application

Prediction will be displayed here:

Nitrogen content in soil:

Phosphorous content in soil:

Potassium content in soil:

Temparature in °C:

Humidity in %:

PH value of the soil (between 3.5-9):

Rainfall in mm:

Fig 16 : Web-based Application

Crop recommendation application

Prediction will be displayed here:

Recommended crop: Mango
Based on the parameters entered, Mango is the suggested crop.

Nitrogen content in soil:

Phosphorous content in soil:

Potassium content in soil:

Temparature in °C:

Humidity in %:

PH value of the soil (between 3.5-9):

Rainfall in mm:




Fig 17 : Web- based Application

Conclusion

The comparative study of three different supervised machine learning models (KNN, Decision Tree, and Random Forest) is done to predict the best-suited crop for the particular land that can help farmers to grow crops more efficiently. In completion, we concluded that

the crop prediction dataset showed the best accuracy with Random Forest Classifier with %. In contrast, has the lowest accuracy among the three with %, and the accuracy of Decision Tree Classifier is in between KNN and Random Forest Classifier.

Future Prospects of the Project

In the future, we will collect new data from the fields to get a clear image of the soil, and integrate other machine learning algorithms with deep learning algorithms such as ANN and CNN to classify more. More factors like wind, altitude level, etc. can be added that can alter the yield of a crop. Further progress can be made to make the prediction real time allowing the model to consider the factors that were not present initially, for example, fertilizer, pesticide, etc. Data can be collected for the same crops with different biological species to figure out which species flourishes in which environment, making the job of the farmer much easier. According to our analysis the model will give more accuracy as the data points increase, so to get better accuracy model data points can be increased. Our system can be integrated with a messaging module so that registered farmers can get the notification of the prediction directly to their registered mobile numbers.

References

- [1] Dahikar S and Rode S V 2014 Agricultural crop yield prediction using artificial neural network approach International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering vol 2 Issue 1 pp 683-6.
- [2] Suresh A, Ganesh P and Ramalatha M 2018 Prediction of major crop yields of Tamil Nadu using K-means and Modified KNN 2018 3rd International Conference on Communication and Electronics Systems (ICCES) pp 88-93 Doi: 10.1109/CESYS.2018.8723956.
- [3] Medar R, Rajpurohit V S and Shweta S 2019 Crop yield prediction using machine learning techniques IEEE 5th International Conference for Convergence in Technology (I2CT) pp 1-5 Doi: 10.1109/I2CT45611.2019.9033611.
- [4] Nishant P S, Venkat P S, Avinash B L and Jabber B 2020 Crop yield prediction based on

Indian agriculture using machine learning 2020 International Conference for Emerging Technology (INCET) pp 1-4 doi: 10.1109/INCET49848.2020.9154036.

[5] Kalimuthu M, Vaishnavi P and Kishore M 2020 Crop prediction using machine learning 2020 Third International Conference on Smart Systems and Inventive Technology (ICCSIT) pp 926-32 doi: 10.1109/ICSSIT48917.2020.9214190.

[6] Pande S M, Ramesh P K, Anmol A, Aishwarya B R, Rohilla K and Shaurya K 2021 Crop recommender system using machine learning approach 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) pp 1066-71 doi: 10.1109/ICCMC51019.2021.9418351.

[7] Ranjani J, V.K.G. Kalaiselvi, A. Sheela, Deepika Sree D, Janaki G 2021 Crop Yield Prediction using Machine Learning Algorithm 2021 4th International Conference on Computing and Communications Technologies (ICCCT) doi: 10.1109/ICCCT53315.2021.9711853

[8] Bharath S, Yeshwanth S, Yashas B L and Vidyaranya R Javalagi 2020 Comparative Analysis of Machine Learning Algorithms in The Study of Crop and Crop yield Prediction International Journal of Engineering Research & Technology (IJERT) NCETESFT – 2020 vol 8 Issue 14.

[9] Mahendra N, Vishwakarma D, Nischitha K, Ashwini and Manjuraju M. R 2020 Crop prediction using machine learning approaches, International Journal of Engineering Research & Technology (IJERT) vol 9 Issue 8 (August 2020).

[10] Gulati P and Jha S K 2020 Efficient crop yield prediction in India using machine learning techniques International Journal of Engineering Research & Technology (IJERT) ENCADEMS – 2020 vol 8 Issue 10.

[11] Gupta A, Nagda D, Nikhare P, Sandbhor A, 2021, Smart crop prediction using IoT and machine learning International Journal of Engineering Research & Technology (IJERT) NTASU – 2020 vol 9 Issue 3.

[12] Janmejy Pant, R.P. Pant, Manoj Kumar Singh, Devesh Pratap Singh, Himanshu Pant Analysis of agricultural crop yield prediction using statistical techniques of machine learning 2021 Volume 46, Part 20
<https://www.sciencedirect.com/science/article/pii/S221478532101052X#section-cited-by>

A. Base Papers

- I. Madhuri Shripathi Rao, Arushi Singh, N.V. Subba Reddy, Dinesh U Acharya Crop

prediction using machine learning Department of Computer Science and Engineering,
Manipal Institute of Technology, Manipal, 576104, Udupi district Karnataka, India
<https://iopscience.iop.org/article/10.1088/1742-6596/2161/1/012033/pdf>

- II. Tandzi Ngoune Liliane, Charles Shelton Mutengwa Factors Affecting Yield of Crops
DOI:10.5772/intechopen.90672
https://www.researchgate.net/publication/342994002_Factors_Affecting_Yield_of_Crops

B. Plagiarism Report

The screenshot shows a web browser window with the URL seomegatools.com/plagiarism-checker. The page title is "Plagiarism Checker" and it includes the instruction "Paste (Ctrl + V) your article below then click Check for Plagiarism!". The report shows "100% Unique Content". A table lists 15 strings from the document, each with a "Good" uniqueness status. To the right, there is a sidebar titled "Our Popular SEO Tools" with links to various tools like CSS Minifier, Check Moz Rank, Keywords Suggestion, Google Pagespeed Insights, HTML Compressor, Image Optimizer, JS minifier, Social Status Checker, English Grammar Checker, and XML Sitemap Generator. The footer of the page indicates "We have over 60 Free SEO tools" and "This site is Proudly hosted in Australia".

Plagiarism Checker
Paste (Ctrl + V) your article below then click Check for Plagiarism!

100% Unique Content

#	String	Uniqueness
1	Prediction of Suitable Crop Using Machine Learning	Good
2	School of Computer Engineering and Technology	Good
3	University, Kothrud,	Good
4	ne 411 038, Maharashtra - India	Good
5	SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY	Good
6	rtify that, Shranay Shahane Sakshi Dongre Nihaal Shetty	Good
7	ande] [Dr. Vrushali Kulkarni]	Good
8	University, Pune MIT World Peace University, Pune	Good
9	University, Pune	Good
10	Internal Examiner:	Good
11	ngineering, Dr. Vishwanath Karad MIT World Peace University,	Good
12	University, Pune for providing us the opportunity and	Good
13	ject guide, Dr. Himangi Pande for guiding and helping	Good
14	ne Project. Her guidance helped us to bring the project	Good
15	countries. Modern agriculture is an ever-expanding	Good

Our Popular SEO Tools

- [CSS Minifier](#)
- [Check Moz Rank](#)
- [Keywords Suggestion](#)
- [Google Pagespeed Insights](#)
- [HTML Compressor](#)
- [Image Optimizer](#)
- [JS minifier](#)
- [Social Status Checker](#)
- [English Grammar Checker](#)
- [XML Sitemap Generator](#)

We have over 60 Free SEO tools
This site is Proudly hosted in Australia