Big Data Analysis

(Term Project)

Indian Institute of Technology
Kharagpur, West Bengal, India, 721302

Group Members

Ajay Jarwal 20MA20003 **Batchu Sri Sai Sanjana** 20MA20015 **Bussa Varshith** 20MA20016

G Abhinav Raj 20MA20022 **Neeraj Gartia** 20MA20038

Parth Patadia 20MA20040 **Soumya Ranjan Mohapatra** 20MA20056 **Sumit Patel** 20MA20058

Ved Prakash 20MA20062 **Aakash Sadhwani** 20MA20076

Project Topic:

Maximizing Farmer's Profit based on Crop Production

Contents

1 Introduction **3**

2 Problem formulation **3**

3 Data **4**
3.1 Dataset Collection 4
3.2 Collecting Environment Factors 4

4 Methods **5**
4.1 Random Forest (RF) 6
4.2 Decision Tree (DT) 6
4.3 Naive Bayes (NB) 7
4.4 XGBoost Classifier 8
4.5 LightGBM 8
4.6 Crop Recommendation 9
4.7 Cross Validation 9

5 Results **10**

6 References **11**

1 Introduction

Agriculture has an extensive history in India. Recently, India is ranked second in the farm output worldwide . Agriculture-related industries such as forestry and fisheries contributed for 16.6% of 2009 GDP and around 50% of the total workforce. Agriculture's monetary contribution to India's GDP is decreasing . The crop yield is the significant factor contributing in agricultural monetary. The crop yield depends on multiple factors such as climatic, geographic, organic, and financial elements. It is difficult for farmers to decide when and which crops to plant because of fluctuating market prices. Citing to Wikipedia figures India's suicide rate ranges from 1.4-1.8% per 100,000 populations, over the last 10 years. Farmers are unaware of which crop to grow, and what is the right time and place to start due to uncertainty in climatic conditions. The usage of various fertilizers is also uncertain due to changes in seasonal climatic conditions and basic assets such as soil, water, and air. In this scenario, the crop yield rate is steadily declining. The solution to the problem is to provide a smart user-friendly recommender system to the farmers.

The crop yield prediction is a significant problem in the agriculture sector. Every farmer tries to know crop yield and whether it meets their expectations , thereby evaluating the previous experience of the farmer on the specific crop predict the yield . Agriculture yields rely primarily on weather conditions, pests, and preparation of harvesting operations. Accurate information on crop history is critical for making decisions on agriculture risk management .

In this report, we have proposed a model that addresses these issues. The novelty of the proposed system is to guide the farmers to maximize the crop yield as well as suggest the most profitable crop for the specific region. The proposed model provides crop selection based on economic and environmental conditions, and benefit to maximize the crop yield that will subsequently help to meet the increasing demand for the country's food supplies. The proposed model predicts the crop yield by studying factors such as rainfall, temperature, area, season, soil type etc.

2 Problem formulation

The potential profit that a farmer can generate in an upcoming season is subject to various factors that influence the production of a particular crop in their region. Among these factors, the weather forecast plays a crucial role in determining the yield of a crop. Additionally, other elements such as soil quality, water availability also impact crop growth and yield. By carefully considering these factors and utilizing weather forecast information, I want to build a model for farmers so that they can make maximum profit by cropping a suitable crop. Various approaches have been proposed for this model, including Decision Tree, Gaussian Naive Bayes, Logistic Regression, Random Forest, XGBoost, LightBGM and Cross Validation.

Archana, K. et al. in this paper developed a system using a voting-based ensemble classifier to recommend the most appropriate crop, fertilizer, and to predict crop yield. The proposed system recommends a suitable crop for cultivation based on Nitrogen(N), Potassium(K), Phosphorus(P), soil type, soil texture, land type, pH, electric conductivity, and temperature parameters. The proposed system also suggest fertilizers to farmers for recommended crop and predict crop yield. The author also developed a crop rotation module, in which alternative crops except predicted crops were suggested to farmers based on crop cultivation seasons.

Chaudhari, A. et al. in this paper proposed a platform for crop recommendation and its optimal seed price using shopbot. A dataset with the following parameters - crop, rainfall, temperature, season, year, production, area, and the location used for crop recommendation. Dataset was normalized and split into a 7:3 ratio.data mining techniques were applied to the dataset, and there accuracy score was calculated. A suitable crop is recommended using the classification algorithm. In their work, for optimal pricing of seeds, web scraping was performed using the selenium framework.

Kelvin Tom Thomas et al. in his work proposed a recommendation system, which recommends the most suitable crop based on soil parameters to the farmers. The different machine learning techniques like Decision Tree(DT), K-Nearest Neighbor(KNN), KNN with cross- validation, Naive Bayes were used to train a model. A dataset with the following soil parameters - soil NPK values, soil pH values, and most suitable crop was used for prediction. The author found that KNN with cross-validation Based on the record present in their dataset value for 'k' has taken as 10. The proposed model predicts the most suitable crop by taking soil NPK and soil pH values.

Lavanya B et al. in this paper presented a crop prediction system using data analytics techniques in the form of an android application. The dataset used in this study includes the following parameters: weather, soil type, soil pH, previous three harvests, fertilizers, season, and market demand. Initially proposed system take input from farmers like details of previous three harvest, availability of water and the data of weather conditions, temperature and soil condition are taken based on users location In their system, based on collected data from users and the dataset available, a suitable crop for cultivation was predicted by applying machine learning techniques.

In conclusion, the problem of crop recommendation has been extensively studied, and various approaches have been proposed to develop accurate and efficient crop recommendation systems. By utilizing machine learning and data analytics techniques, crop recommendation systems can provide valuable insights to farmers, leading to better crop yields and increased profitability

3 Data

3.1 Dataset Collection

In order to achieve a healthy harvest of plants, it is essential to provide optimal conditions such as temperature, humidity, soil pH, rainfall, and nutrient levels (NPK) to support plant growth. These conditions may vary depending on the plant species being cultivated. To ensure accurate recommendations for crop cultivation, an initial dataset was collected from reliable sources such as indistats, data.gov.in, and research papers. This dataset was then used to train a crop recommendation model, with the aim of increasing accuracy in recommending suitable crops for specific conditions.

N	P	K	temperature	humidity	ph	rainfall	label
90	42	43	20.879744	82.00274	6.502985	202.9355	rice
6	62	22	20.530527	18.09224	5.824091	120.4509	kidneybeans
26	73	21	31.331708	57.97429	4.946264	161.782	pigeonpeas
22	43	24	25.42517	53.22083	4.523636	46.19375	mothbeans
32	57	22	28.689985	87.50437	6.769416	44.56598	mungbean
29	16	36	19.810694	88.92944	5.740338	102.8601	pomegranate

Figure 1: Sample dataset

3.2 Collecting Environment Factors

Temperature: Temperature plays a crucial role in plant growth and development. Different plants have specific temperature requirements for optimal growth. Deviations from the ideal temperature range can result in reduced growth and lower yields. The initial dataset collected from reliable

sources provided information on the temperature requirements of various plant species, which was then used to train the crop recommendation model. By considering the temperature data, the model can accurately recommend crops that are suitable for the prevailing temperature conditions in a given region, ensuring optimal plant growth.

: Humidity, or the amount of moisture in the air, is another important factor that affects plant growth. Different plant species have varying humidity requirements for optimal growth. High humidity levels can promote the growth of pests and diseases, while low humidity can result in water stress for plants. The initial dataset collected from reliable sources included information on the humidity requirements of different plant species, which was used to train the crop recommendation model. By considering the humidity data, the model can provide accurate recommendations for crops that thrive in the prevailing humidity conditions of a specific region.

Soil pH: Soil pH, or the acidity/alkalinity of the soil, is a critical factor that affects plant growth. Different plants have specific soil pH requirements for optimal growth. Deviations from the ideal soil pH range can result in nutrient deficiencies or toxicities, which can negatively impact plant health. The initial dataset collected from reliable sources provided information on the soil pH requirements of various plant species, which was used to train the crop recommendation model. By considering the soil pH data, the model can recommend crops that are suitable for the prevailing soil pH conditions in a given region, ensuring optimal nutrient uptake and plant growth.

Rainfall: Rainfall, or the amount of precipitation a region receives, is a vital factor that affects plant growth. Different plant species have varying requirements for rainfall, depending on their water needs and tolerance to water stress. Insufficient or excessive rainfall can result in reduced yields and crop failures. The initial dataset collected from reliable sources included information on the rainfall requirements of various plant species, which was used to train the crop recommendation model. By considering the rainfall data, the model can accurately recommend crops that are suitable for the prevailing rainfall conditions in a given region, ensuring optimal water availability for plant growth.

NPK Values: NPK values refer to the levels of essential nutrients, namely nitrogen (N), phosphorus (P), and potassium (K), in the soil. These nutrients are crucial for plant growth and development, and different plants have specific requirements for NPK levels. Imbalances in NPK levels can result in nutrient deficiencies or toxicities, which can adversely affect plant health and productivity. The initial dataset collected from reliable sources provided information on the optimal NPK levels for various plant species, which was used to train the crop recommendation model. By considering the NPK data, the model can recommend crops that are suitable for the prevailing NPK levels in a given region, ensuring optimal nutrient availability for plant growth..

4 Methods

In the first step, the crop dataset was collected. Two distinct datasets were collected and merged based on similar crop labels. The prepared dataset contains the following parameters - temperature, humidity, pH, rainfall, Nitrogen(N), Phoshorus(P), Potassium(K), and crop label. After finalizing data, in the next step, data is preprocessed. Data Pre-processing is the most important task because the collected dataset mostly contains missing values, which affect the accuracy of the model. Data were preprocessed and then divided into two datasets: training and testing. 80% of data was used for training, while 20% was kept for testing model performance. The five machine learning classification algorithms like the Decision Tree, Random Forest, Gaussian Naive Bayes, XGBoost Classifier, LightGBM were used for training the prediction model. After training performance of these four techniques was compared. The final model has been built using XGBoost Classifier. Following the model training, a predictive model was developed, which will predict a suitable crop for cultivation based on given data.

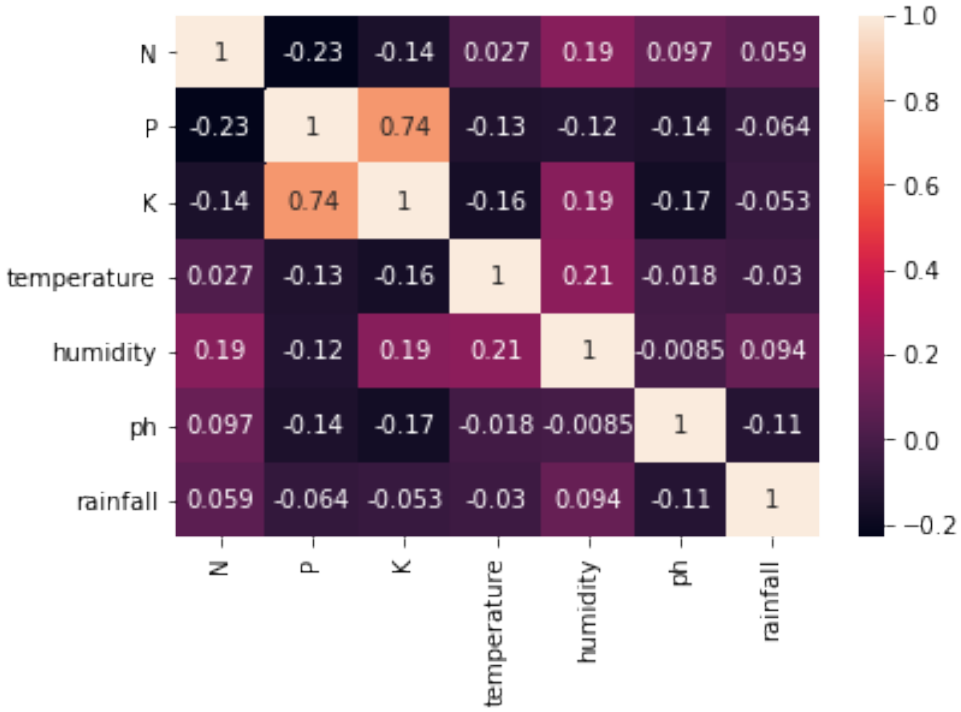

Figure 2: Heat Map

4.1 Random Forest (RF)

Random Forest is also one of the most popular supervised machine learning techniques used for solving both regression and classification problems. Random forest is a commonly-used machine learning algorithm which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fuelled its adoption, as it handles both classification and regression problems. Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems. The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged, and for a classification task, a majority vote—i.e. the most frequent categorical variable—will yield the predicted class. Finally, the oob sample is then used for cross-validation, finalizing that prediction.

4.2 Decision Tree (DT)

Decision trees are commonly used for both regression and classification problems. The decision tree algorithm creates a tree-like model of decisions and their possible consequences, where the leaves represent the classes or the regression output. To determine which attribute to choose as the root node and subsequent nodes, different attribute selection criteria are used, such as information gain, Gini index, and others. The Gini index is one of the popular methods for selecting attributes in decision trees. The formula for calculating the Gini index is as follows:

where p1, p2, ..., pk are the probabilities of each class in a given node. The Gini index measures

$$\text{Gini}(D) = 1 - \sum_{i=0}^{n} P_i^2$$

the impurity or the degree of classification error at a node. The lower the Gini index, the better the attribute is for splitting the data into pure classes. Once the decision tree is built, it can be used to make predictions for new data by traversing the tree from the root node to the leaf node that corresponds to the predicted class or regression output.

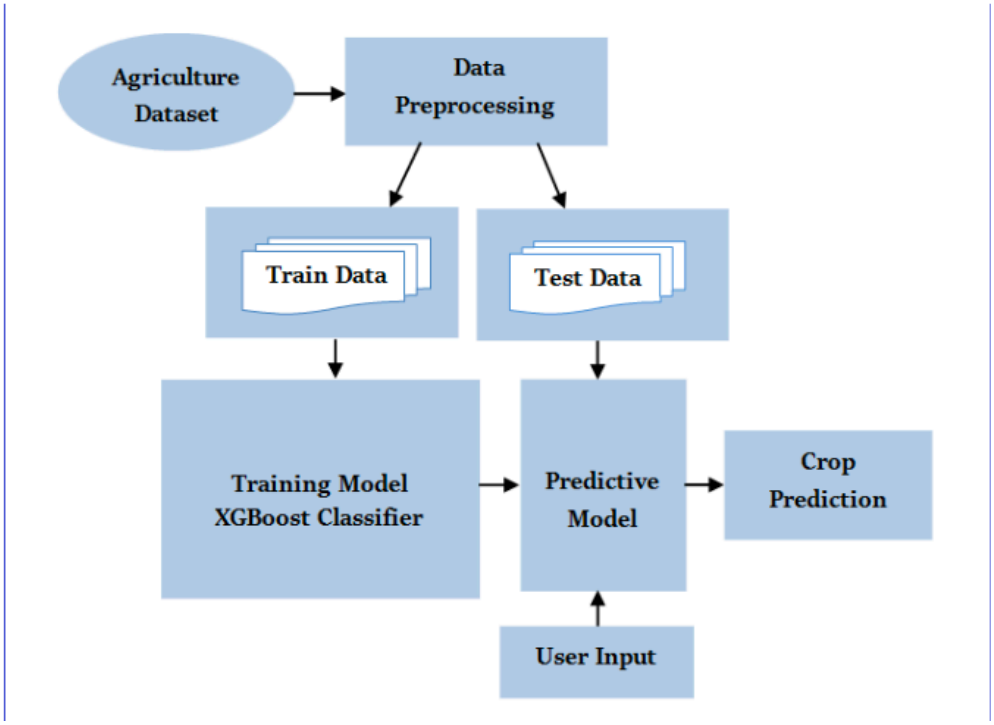

Figure 3: Machine Learning (ML) Model

4.3 Naive Bayes (NB)

Naive Bayes is a machine learning algorithm that is commonly used for classification tasks. It is based on Bayes' theorem, which describes the probability of an event based on prior knowledge of conditions that might be related to the event. In Naive Bayes, the algorithm assumes that the features used to classify an instance are independent of each other, which is often not the case in real-world scenarios. Despite this simplification, Naive Bayes has been shown to be very effective in many applications, especially when there are a large number of features. The algorithm works by calculating the probabilities of each possible class based on the values of the features for a given instance. It then selects the class with the highest probability as the predicted class for the instance. There are several variants of Naive Bayes, including the Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes. Each variant is suited to different types of data and classification problems.

$$P(x \mid y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

4.4 XGBoost Classifier

XGBoost, stands for extreme gradient boosting, is a powerful decision tree-based boosting algorithm that is widely used in machine learning for solving various problems including classification, regression, and prediction. It is implemented on top of the gradient boosting framework, which is an ensemble technique that combines multiple weak models (e.g., decision trees) to create a strong model.

One of the key advantages of XGBoost is its speed and accuracy. It is known for its high computational efficiency, making it suitable for handling large datasets. The algorithm is designed to be highly optimized and uses a parallelization concept to process data in parallel, further enhancing its performance. This makes XGBoost a popular choice for many real-world applications where both speed and accuracy are crucial.

In addition to its speed and accuracy, XGBoost also addresses the issue of overfitting, which is a common challenge in machine learning. Overfitting occurs when a model learns to perform well on the training data but fails to generalize well to unseen data. XGBoost employs regularization techniques, such as tree pruning, which prevents trees from growing beyond a certain size, thereby reducing the risk of overfitting. This helps in creating a more robust and generalizable model.

Moreover, XGBoost performs well on small to medium-sized datasets. It is capable of handling datasets with a limited number of samples and features, making it suitable for applications where data availability is limited.

4.5 LightGBM

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework developed by Microsoft that uses tree-based learning algorithms. It is designed to be efficient and scalable, and can handle large datasets with a high number of features. LightGBM uses a gradient-based One-Side Sampling (OSS) technique to construct trees and also employs Histogram-based algorithms to compute the splitting gain efficiently. Additionally, it supports GPU acceleration to further improve training speed. LightGBM is often used for tasks such as classification and regression, and has been shown to achieve state-of-the-art results on various benchmark datasets.

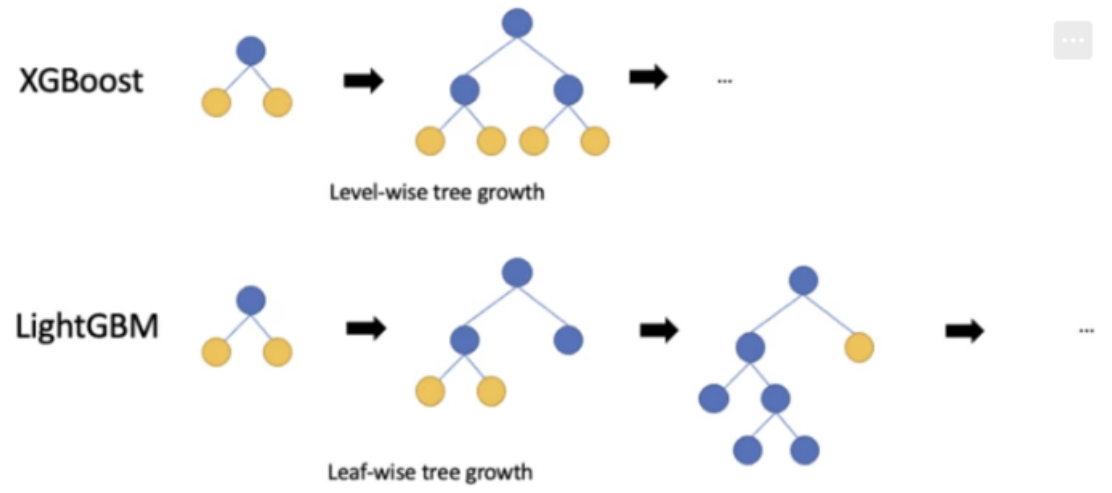

Figure 4: Tree growth of XGBoost and LightGBM

4.6 Crop Recommendation

Our advanced predictive model, powered by cutting-edge machine learning techniques, revolutionizes crop recommendations for the Pune district. By incorporating soil parameters, crop season, and weather data, our model provides accurate and reliable predictions on the best crop for cultivation. Users simply input location, soil moisture, soil type, and crop season information, and our system employs the XGBoost classifier to deliver optimal crop recommendations. This user-friendly solution aims to improve crop yields and maximize profits for farmers in Pune district, providing a practical and efficient approach to crop selection for sustainable agriculture.

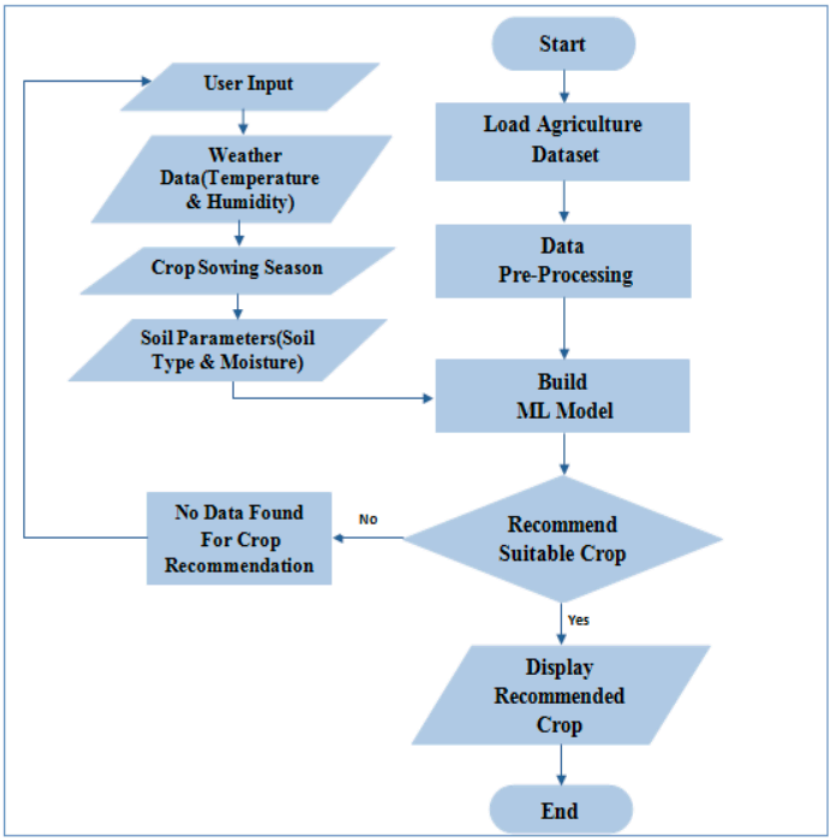

Figure 5: Flowchart of Crop Recommendation System

4.7 Cross Validation

Cross-validation is a technique used in machine learning and statistics to evaluate the performance of a model on an independent dataset. The basic idea behind cross-validation is to divide the dataset into multiple subsets or ”folds”, where one fold is used for testing the model and the remaining folds are used for training the model. This process is repeated multiple times, with each fold used once for testing and the rest for training, and the results are averaged across all the folds to obtain an estimate of the model’s performance.

The most commonly used type of cross-validation is k-fold cross-validation, where the dataset is divided into k equal sized folds, with one fold used for testing and the remaining k-1 folds used for training. This process is repeated k times, with each fold used once for testing, and the results are averaged across all the folds to obtain a more reliable estimate of the model’s performance.

Cross-validation is an important tool for evaluating the performance of a model, especially when the dataset is small or when the model is prone to overfitting. It can also be used to compare the performance of different models or to tune the hyperparameters of a model.

5 Results

Seven classification algorithms such as Decision Tree, Logistic regression, SVM, Random Forest, Naive Bayes, and XGBoost classifier, LightGBM were used for crop prediction. The accuracy of the Decision Tree classifier without hyperparameter tuning was 90%, and Random Forest algorithm with hyper parameter tuning was 99% on test data. The accuracy of the Naive Bayes, SVM, logistic regression algorithm was 98% , 97%, 95% on test data respectively. The accuracy of XGBoost Classifier with hyperparameter tuning was 99.3% on test data. As the XGBoost classifier gave the highest accuracy as compared to other algorithms, hence the final model was developed using the XGBoost classifier. Same is depicted in the graph below.

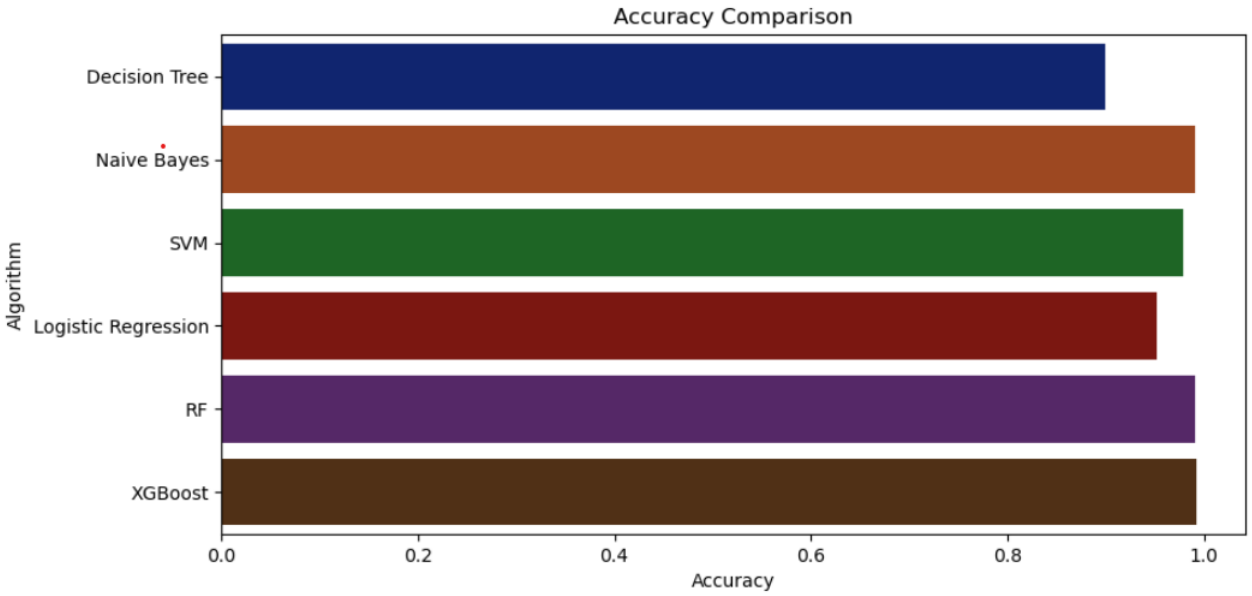

Figure 6: Accuracy Comparison among different model

Agriculture has always been an important activity in India because it's not only a source of food but also a source of income for some people. As a result, in order to boost agricultural production, it must be done efficiently. There are various factors that affect crop growth, so choosing the right crop will always increase crop productivity. A crop recommendation system is proposed to help farmers in selecting the best crop for cultivation based on their location, soil factors, and weather parameters. Various machine learning approaches were used, such as decision tree classification, random forest classification, Naive Bayes classifier, and XGBoost classifier. Out of them, the XGBoost classifier gives the best results with a 99.3 accuracy, hence the final model was developed using the XGBoost classifier. In the future, for crop recommendation, more soil parameters are going to be considered and data of more districts will be taken into consideration for more precise results.

6 References

[1] Archana, K., Saranya, K. G. (2020). Crop Yield Prediction, Forecasting and Fertilizer Recommendation using Voting Based Ensemble Classifier. SSRG Int. J. Comput. Sci. Eng, 7.
[2] Chaudhari, A., Beldar, M., Dichwalkar, R., Dholay, S. (2020, September). Crop Recommendation and its Optimal Pricing using ShopBot. In 2020 International Conference on Smart Electronics and Communication (ICOSEC) (pp. 36-41). IEEE.
[3] Kevin Tom Thomas, Varsha S, Merin Mary Saji, Lisha Varghese, Er. Jinu Thomas. Crop Prediction Using Machine Learning. International Journal of Future Generation Communication and Networking. Vol. 13, No. 3, (2020), pp. 1896–1901.
[4] Lavanya, B., Nisarga, B., Meghana, B. S., Mythresh, A. CROP PREDITION USING MACHINE LEARNING.
[5] https://data.desagri.gov.in/website/crops-apy-report-web
[6] https://datasource.kapsarc.org/explore/dataset/district-wise-rainfall-data-for-india-2014/table/
[7] https://community.data.gov.in/tag/crop/