# Enhancing Crops Production Based on Environmental Status Using Machine Learning Techniques

by

Shiyam Talukder
16101243
Habiba Jannat
16101191
Sukanta Saha
20141019
Katha Sengupta
16101280

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
April 2020

# Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

| | |
|---|---|
| Shiyam Talukder | Katha Sengupta |
| 16101243 | 16101280 |

| | |
|---|---|
| Habiba Jannat | Sukanta Saha |
| 16101191 | 20141019 |

# Approval

The thesis titled "Enhancing Crops Production Based on Environmental Status Using Machine Learning Techniques" submitted by

1. Shiyam Talukder (16101243)

2. Habiba Jannat (16101191)

3. Sukanta Saha (20141019)

4. Katha Sengupta (16101280)

Of Spring, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on April 07, 2020.

**Examining Committee:**

Supervisor:
(Member)

—————————————————
Muhammad Iqbal Hossain
Assistant Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

—————————————————
Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

—————————————————
Dr. Mahbub Alam Majumdar
Professorand Chairman, Interim Dean, School of Science
Department of Computer Science and Engineering
BRAC University

# Abstract

Bangladesh is an agricultural country. As the economy is based on agriculture highly, there should be progress in this sector. To make progress in agriculture the productivity must be increased. These days, productivity is low due to various factors. One of them is not finding suitable crops for a particular land. In this way, the crops are not produced at the maximum amount. Hence, productivity of agriculture depends on multiple parameters on the basis of location. The suitable crop for a particular location is necessary for agriculture to bring the most productivity. Here we have designed a model that predicts productivity with given parameters, and also recommends the suitable crop based on those parameters. In terms of Machine Learning for the prediction and the recommendation, we have applied multiple algorithms like k-nearest neighbor, support vector machines, random forest, naïve Bayes' classifier and logistic regression, collaborative filtering and fuzzy K-Nearest neighbor. After training the dataset and applying algorithms, for prediction we have made a comparison by analyzing the precision. On the other hand, for recommendation we have used collaborative filtering system and fuzzy k-nearest neighbor. These algorithms are mainly used to take users data as input and test with the trained data that is already in the system and will filter out the best 5 crops as output.


**Keywords:** Agricultural Productivity; Recommendation; Prediction; Precision; K-Nearest Neighbor; Collaborative Filtering; Fuzzy K-Nearest Neighbor

# Acknowledgement

First and foremost, we are grateful to God the almighty, to Whom the wisdom and power belongs, for His blessings and love to His creations.
Our appreciation goes to our thesis supervisor Dr. Muhammad Iqbal Hossain for granting a wonderful opportunity to work with him and to support and guide us in this thesis.
We are also indebted to everyone who helped us to proceed in our thesis work.

# Table of Contents

# List of Figures

# List of Tables

# Nomenclature

Abbreviation list below will be used later within the body of the document.

$AI$     Artificial Intelligence

$CF$     Collaborative Filtering

$fKNN$   fuzzy k-nearest neighbors

$KNN$   K-Nearest Neighbor

$LR$     Logistic Regression

$ML$     Machine Learning

$MMH$   Maximum Marginal Hyperplane

$SVM$   Support Vector Machines

# Chapter 1

# Introduction

Ever since the very beginning of human civilization, agriculture has been the root of the development. As food is one of the basic human needs, agriculture has been the source of the supply of it. On the other hand, agriculture affects the economy by providing raw material and food to non-agricultural sectors. It has a good support of employment amongst rural dwellers, providing them with purchasing power. The dominance of agriculture can be observed from non-developing countries to developed countries. The developed countries have a history of good agricultural productivity that led to the enhancement of the economy. The agricultural activities give the economy some power so that it can enrich itself for betterment and upgrades the economy. Thus, agriculture has given a proper structure to the source of investment. As for developing countries like Bangladesh, most of the industries here are based on agricultural products including pharmaceutical and food processing. In this way, it has always played a strong backbone for a nation's flourishment. Back in the Neolithic revolution, agriculture empowered the growth of the human population more than hunting or gathering. Thus, agriculture is the root of human survival.

Even with the transformation of the agrarian societies into industrial societies, as it mentioned before, countries like Bangladesh depend on agriculture as it plays a vital role on macro-economic objectives like reducing unemployment or poverty, achieving food security. The economy here has an increased demand of other raw materials that comes from agriculture to support the population. But the drastic increase of population with the increasing demand of staple food and the decreasing arable land has a negative effect on the economy. The agricultural GDP is declining day by day and it clearly shows how we are not getting enough efficiency to produce the crops. The efficiency that can be measured by making choices by the selection of suitable crops for suitable lands.

Since the foundation of agrarianism we have noticed how much agriculture is dependent on multiple parameters of the environment and surroundings. Even though agriculture has been reducing hunger and other basic human needs, the sector of agriculture has been underdeveloped. There is a lot of failure, lack of production that takes place for the need of making rational choices. If there has been rational choice in agriculture, then the productivity would be high and there would be more profit from it as the proper choices of agriculture would reduce the cost and money. Hence, there should be a proper study for the betterment and benefits in it so that cultivators get more results than the effort. The failure is very unexpected, the

investment in agriculture is very crucial and struggling for developing countries. If there are predictions that cannot take place, then there can be more losses than profits. As agriculture depends on nature, the growth of the crops depends on various components of the climate such as rainfall, humidity, temperature and also it is dependent on multiple seasons. Thus the determination of choosing the suitable crop amongst the alternatives for a particular area can be very challenging to decide by the cultivators. It is high time to achieve maximum production of the crops to reduce hunger and poverty as agriculture deals with food security and as well as growth in the economy.

## 1.1  Motivation

Agricultural productivity refers to the efficiency of the system, such as the output to input ratio. It is dependent on inputs of different aspects such as climate, weather, season, factors etc. For one slight change of inputs the output varies in a wide range. The productivity of agriculture determines the efficiency of the effort of cultivators in such a way they are benefited as well as the economy. It helps the growth of agriculture as well as it eradicates poverty. the rural portion of the economy relies on agriculture, the productivity would increase employment and upgrade the status of livelihood. The productivity of agriculture would also help the decrease the prices of the food and it stabilizes the supplies. The productivity helps the investors to gain profit and invest more on the production of agriculture and it attracts people to earn a living by farming and moreover as a contribution to achieving food security. People these days are moving away from agriculture and focusing more into industrialization forgetting the impact of agriculture in industries by providing raw materials and also how it helps the environment to be safe from pollution. If people are benefited from agriculture more than industrialization, then they would be conscious enough to make the proper choices. Hence, increasing the production of the crops can be the solution. Increasing the production of the crops are only possible if there is proper utilization of resources. Farmers need to be well aware of the proper choices. Yielding crops need fuel and other sources of energy, labor cost, time. If the maximum amount of production cannot be achieved, then it is a waste to the government. So, there are a lot of problems if productivity is not gained. In this way we have thought of a way of solution for such problems. Hence, we highly believe that few machine learning algorithms could be very beneficial to determine the suitable crops and also predict the productivity of any particular land. It is very challenging to study a huge number of dataset for prediction and suitable crops. We have taken effective approaches for such problems with application and study of the algorithms. We have noticed the traditional approaches of predicting maximum yield of crops, and had the comparison with our studies for a better result. With the study of past data, it can help to predict future results so that the choices are rational. We have thought of our ways in which cultivators get to choose the best alternative paths where they get the maximum benefits. Therefore, we propose a system that can predict productivity as well as recommend the suitable crop through the potentials of the algorithm of machine learning.

## 1.2 Problem Statement

Bangladesh has been importing food to meet up the demand. For instance, the annual demand for the staple food rice has now exceeded the supply by the ratio of -0.28% [9]. If a country is content with food, in the time of scarcity or breakage of communication system amongst other countries to imports, the country will not face famine. As in 2020, due to the Covid-19 outbreak the communication system has fallen and it has put the demand of food in risk through importation. As it is a global pandemic, the labor would be scarce due to lockdown, the supply of agricultural equipment would fall due to the situation. For instance, in Europe due to the massive outbreak of the Covid-19 virus, the border restriction has caused the shortage of laborers. For countries like Italy where the outbreak took a massive toll of death and affected people where a large portion of the food production is produced by seasonal workers from neighboring countries. As the workforce has fallen into travel restriction for the lockdown, it is very challenging to achieve food security in this time. Even for Bangladesh, due to the lockdown the productivity of the whole population has fallen, the transport system of all over the country has been decreased. In this time if the massive outbreak like Europe happens in Bangladesh, it will cause famine as it happened before in global crisis time such as the famine of 1943 during the World War II due to supply shortage. Thus, if the productivity is increased then there is agricultural growth. In this way, if there is plenty of food supply then the country can be content when the population has fallen into the risk of pandemic diseases.

The contentment can be achieved by producing a maximum amount of crops, in that way in the time of scarcity it can be a great relief. Unfortunately, it is very challenging to yield abundant crops in limited arable land. Due to the climate change the productivity of land has been changed. Previously Bangladesh has been self-dependent for ensuring staple food to feed the population. In these days, due to the population increase, the agricultural land has been 70.6% (with 59.65% arable land and only 6.38% permanent cropland) in 2016 whereas in 2005 it has been 71.5%[4]. Flood, heavy rainfall, less cold winter and every year with record amounts of temperature in summer has a good deal of effect in cultivable lands. Every year, the temperature in summer is increasing at the rate of 1 degree Celsius. By the year 2050, the increasing temperature will be effective on GDP by decreasing 6.7%[22].The amount of resources is very limited in our economy. Factors of production is very scarce. The declining arable land, high price of labor, lack of capital that mostly comes from money lenders makes the investment very struggling. In this way to gain the productivity the resultant crops are measured with minimum effort but yet maximum gain. The productivity matters to minimize the losses, if there is no prediction of productivity then without proper calculation and depending on the nature it is very problematic to determine the desired result. Consequently, the efforts will go wasted, the borrowed capital that cannot be repaid and also there are equipment that are circulated through the fuels. The fuel source is also very scarce everywhere. Thus, the better the condition of inputs the greater the productivity of agriculture.

The development of a country depends on the economy and our economy depends on agriculture. The importance of agriculture in our economy is very significant to eradicate poverty since 2000 and from 2005 to 2010, agriculture has the ability to

remove 90 percent poverty in rural areas[18]. Apparently, the change is not noticeable but within a few decades. The percentage will decrease more and more. Hence, the production of food will decrease to feed this vast population.

In Bangladesh, the technology is getting advanced as well as agriculture. Still the forecasting and prediction in agriculture has been insufficient. The productivity has been very unpredictable. Most of the cultivators depend on their previous experience to cultivate their land. As they are dependent highly on climate for making decisions for agricultural production so if there is change in climate pattern, they do not know which crop to produce. In these years, the productivity of the crop fluctuates due to the changing parameters of climate. Therefore, they cannot produce the desired maximum amount of crop that could be produced in that area. In a specific area, there is a suitable crop accordingly for the best productivity. In this situation, we propose our approach involving machine learning which will take a particular district, season, rainfall, humidity and temperature as parameters and apply suitable algorithms for prediction and recommendation.

## 1.3 Objectives and Contributions

We have collected our data set from an authentic source of Indian websites and used that dataset to work with our thesis experiment. Firstly, we used the data from the dataset to compare with different machine learning algorithms to find the accuracy in predicting Agricultural Productivity. In our work, in the phase of data pre-processing and we divided the dataset for training and testing. The set for training is implemented in various algorithms due to prediction and recommendation. In Addition, we have made a comparison between the prediction algorithms with precision.

Recommendation of crop is the main domain of agriculture. As we know in most developing countries, the majority of the population live in rural areas and they earn their living only from the agricultural sector. Agriculture accounts for a significant part of the economic activity in the developing world. Thus, increasing agricultural productivity would contribute to economic development. Here, the mistake the farmers make is choosing the wrong crop for their cultivation despite different needs of the environmental status. For this reason, they could not meet the desired requirements of the productivity of that crop and this is how the economy collapsed. Not only this, there is a huge demand for varieties of crops in our country due to the increased population. Farmers are unable to meet the demand of the people of our country because of their wrong decision of choosing crops according to their environmental status. At the same time, their earning source condition gets worse and income gets lower which affects the economy of our country. Therefore, in our search we have approached in such way that will not set back the productivity of the agriculture rather it will meet the expectations of the farmers in great ways. In our model, we have used Fuzzy KNN Algorithms, Collaborative Filtering for Recommendation of the crop and naïve Bayes classifier (72%), K-nearest neighbor (KNN) (64%), support vector machine (SVM) (56%), logistic regression (70%), random forest (70%) Algorithms for the precision of the agricultural productivity. It is found that after a comparative result analysis of all the algorithms, the naïve Bayes classifier Algorithm has given the highest precision of 72% for different crops under several parameters among the other Algorithms.

## 1.4  Thesis Orientation

Chapter 1: Introduction where motivation, problem statement, objectives and contributions are discussed

Chapter 2: Background where literature review and related works of our research are viewed and also the algorithms implemented in our models

Chapter 3: Proposed Agricultural Productivity Model, here we have discussed about dataset description where dataset preprocessing and selection of crop trait and how our model of recommendation made are described

Chapter 4: Result analysis, here we will compare the results with algorithms to analyze the accuracy and build the model of recommendation with data visualization

Chapter 5: Conclusion, we will conclude our research based theory with our models that are made and discuss future works to do for further improvement and References

# Chapter 2

# Background

Choosing the suitable crop amongst the alternatives or prediction of productivity is based on the geographic location, the weather report on a specific season of the year. The factors depend highly on the involvement of the growth of the crop. We have approached differently with machine learning algorithms for prediction. The algorithms of machine learning create a mathematical structure from a dataset through training. The wide range use of machine learning inspired us to work on such problems dealing with agriculture. Machine learning techniques train the dataset in such a way that it makes the "machine" think, or take logical attempts to solve real life issues. The impact of machine learning has on large scale operation, in economic activity or even on basic life sciences too. Relating technology with daily life is a challenging attempt, using the information technology outside the device and connecting it with more natural sciences like agriculture. Previously it has worked on many real life agricultural sectors such as for livestock in China the companies like Alibaba and Dekon group used artificial intelligence as it recognizes the health of particular animals through the data of temperature and movements [21]. Another example of machine learning in agriculture is the attempt of the precision farming technique which is applying historical data for specific decisions with implementation in a small area. Precision farming takes the help of software and hardware for the measurement and variability in the crop and while monitoring the crop it can predict the early damages and collect the data in such a way so that it can reduce losses of equipment, time and fuels.

The sense of forecasting refers to the prediction which is the output of an algorithm. The dataset training is based on past events that brings out a particular outcome with satisfied results. As in, for our system the algorithms study the previous data and forecast for the future. to evaluate we have compared the outcomes for the best result. In this way for a new record in this dataset, the system will be able to read and evaluate it for the outcomes. Forecasting has been applied to different aspects of our daily life such as weather, business, market sales. It helps us to make rational decisions in multiple alternatives. In this way we have determined to apply machine learning in the problems that we have mentioned so that the cultivators would take rational decisions for a better outcome.

The practice of recommendation is basically amongst the internet users. In such a way, it shows the preference based on choices to save time and energy. In this way we have divided our approaches to achieve the productivity of agriculture.

## 2.1    Literature Review

Here we have studied the application of different machine learning algorithms in various ways. For instance, In [12], D. S. Zingade et al. used various machine learning techniques to build a system where it takes the data of weather and soil condition and based on it the system provides a solution to predict the most profitable crop for. The system is an Android based application that analyses the data for accurate output. The system combines the data from the repository and weather department and implements machine learning algorithms. Here the algorithm they have applied is multiple linear regression which is a prediction algorithm. The system takes the input by the help of GPS and by that the system provides the profitable crop to the farmer. The previous data is also in the consideration of the system for the most profitable crop to the desired area. In this way the result is found through training and testing and the algorithms consider the input of weather, soil. After this, data process the cultivable crops with the help of previous production and predict through multi-linear regression with weather, soil material and previous production then the system calculate the profit through the parameter of prediction and current crop and it shows the output of maximum profitable crops with proper requirements which works for suggestion. In this way the system works with agricultural productivity helping the cultivators take the decision of proper crop and also gives the suggestion to the them. the system takes the idea of previous production, soil, area into consideration for the best result.

Not only we studied about the algorithms but also we observed the variation of databases. Such as, in [1], S. S. Dahikar and S. V. Rode implemented the ANN algorithm to determine the suitable crop for the specific soil type. As the soil has a lot of parameters such as pH, nitrogen, temperature and the effect of the climate on the soil, they have predicted one particular crop to be yielded in one particular type of soil depending on the parameters. Here they had a comparison of different types of algorithms such as artificial neural network, information fuzzy network and other data mining techniques and with the comparison they have come to a conclusion where they have claimed the Artificial Neural network algorithm is the most suitable to their system. This paper provides a comparison study of different algorithms where there is accuracy measured and the best algorithm is chosen. They have claimed in the algorithm that the Artificial Neural network is good for prediction to their system using the soil, weather and market price as the given parameter.

The accuracy of some algorithms has performed better in some situations. Such as, in [2], Mr. V. Lamba and Dr. V. S. Dhaka discussed the techniques of wheat yield prediction comparing Artificial Neural Network (ANN) with other models such as Multiple Linear Regression (MLR), Logistic regression, Time series model etc. With their research they have claimed the accuracy and efficiency of ANN better than others. They have used different types of models on different aspects. For Multiple Linear regression and logistic regression models they predicted crop yielding through plant characters and pest count. They have implemented the probabilistic model Markov chain to forecast the techniques in agriculture. Moreover, In [16], authors have compared between different AI models for production in the Midwestern region crops in the United States. They have initially selected the summertime for specific crops and applied the model on it. With a comprehensive study they had a

7

comparison and with proper measurements they have proven that the Deep Neural Network is suitable for their study. They have used satellite images, meteorological data, hydrological data, crop yield statistics and cropland for the dataset in such a way so that it can give the desired outcome. The satellite products have been used to extract cropland pixels for corn. In this way they have combined different techniques for an ideal solution. To summarize, Using the DNN model they have performed an optimization process for the accuracy and they have found it has out-performed rest five of them with prediction error around 7.6% for corn and 7.8% for soya bean for which the error is much less than other methods. In [10], the authors developed a model for forecasting agricultural products by a new type of model which is able to perform on both linear and nonlinear forecasting functions. They collected export data of shrimp and chicken of Thailand performed the algorithms on it. Here they used a model combined with several mathematical equations and then SVM was applied and cross- validation process was performed. Their study showed that SVM with cross-validation is accurate forecasting tools than others for their dataset.

For the recommendation part we have also studied the techniques to apply by relating with other problems. For instance, it has always been the part of preferences in other words the recommendation engines are built by similar choices or similar types of users. Hence, we have gotten our ideas and approaches to apply to work on our problems for the recommendation through the related papers. In [17] the authors have developed a framework for E-learning recommender. Here they have proposed a platform where the system has multiple layers including recommendation models and supplementary data like web logs. They have discussed multiple algorithms that can be implemented to build recommendation engines such as content based, collaborative bas r or knowledge based, hybrid filtering recommendation. From this paper we have learned about a few algorithms, especially for knowledge based where the user ratings are insufficient, or hybrid filtering the recommender combines the advantages of several techniques to minimize the problems. Moreover, we also have related our problem with webpage viewing. Such as, in [7] authors have built a recommender system called PRES with object-oriented programming language Java. It has implemented the content-based filtering by vector space model where documents and profiles have been represented as vectors. Based on the profiles and recently visited pages, the documents have been matched. Hence, the user matches with the preferred document through previous interactions and preferences. This system has shown how users find their preferred pages with the algorithm implied.

Several recommendation systems have been implied in the field of agriculture. In [11], authors have applied collaborative filtering that offers recommendation to the users by developing a system with the help of big data for the complexity. They have clustered the data by the characteristic similarities using an agglomerative hierarchical clustering algorithm, then applied Pearson correlation coefficient for the similarity (Collaborative filtering). With this system, farmers buy and sell the products by the application they have designed. Also, in [14], M. Kuanr, B. K. Rath and S. N. Mohanty built a crop recommender system with inference engine and knowledge-based engine together with input variables as temperature, humidity and rainfall. They have collected the data based on questionnaire assessment then created a database to predict the outcome of suitable crop. They applied cosine similarity to find similar farmers then used the inference rules for the recommendation.

They have also proposed pesticides, fertilizers and other seeds.

## 2.2 Algorithms

Machine Learning is the process where firstly things are clearly examined with experience and know the form of those observed data and then making a better decision in future. Here, the main aim for technological devices like computers is to make it learn how to do things by itself rather than taking any help from human beings. In fact, AI is everywhere so we are also interacting with Machine learning everywhere every day. ML is used in Virtual Personal Assistants like Alexa, Smartphones, Traffic prediction using GPS navigation, Transportation Network as Ubers, Pathao, Email Spam and Malware Filtering, Search Engine, Product or weather Prediction, Online Fraud Detection and in health sector such as cancer treatment, making medicine drugs and automated cars and many more. In our research model, we have used Supervised Learning that is mainly used for future prediction using the data given. In case of Supervised Learning, the Classification method is used in order to put the data or the input value into their right class and the value must be discrete. On the other hand, Regression is to predict a value for an input. Whereas in unsupervised learning, the Clustering method is used to find different separate groups within the whole group of data. Below is our algorithms that we have used so far for our models.

### 2.2.1 Algorithms for Precision

**Naïve Bayes Classifier**

It is a supervised learning algorithm which is used to classify data into predefined classes. It is a conditional independent because [3] all attributes are independent given the value of the class variable. Its main uses are in text classification, content arrangement, spam separating, sentiment analysis and many more. It is fast and easy to work with data. Therefore, it is a method for predicting the probability of independent attributes given a class using the Bayes' Theorem. The Bayes' Theorem:

$$P(A|B) = P(B|A)P(A)/P(B) \tag{2.1}$$

Where P(A—B) is the probability that is needed to find the naive classifier
Ultimately, for prediction the formula used is

$$y = argmax \, p(y = C_k) \prod p(x|y = C_k) \tag{2.2}$$

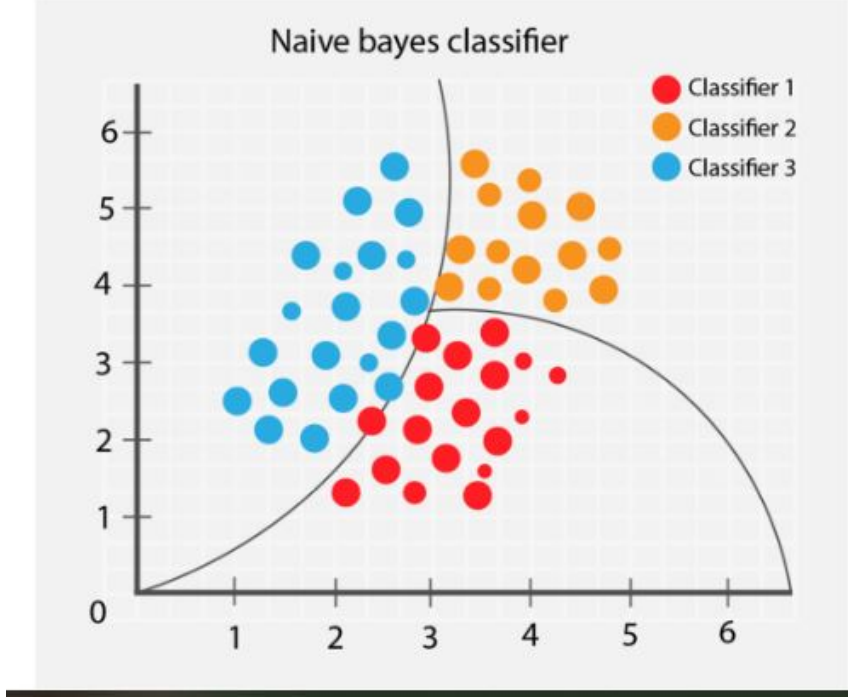The overall workings of the Naïve Bayes' Classifier are shown below in figure.

Figure 2.1: Naïve Bayes' Classifier

**Logistic Regression**

It is the statistical model which is used to estimate the probability of the categorical values and it is also used as a classification model for prediction. It is a classification method for which it gives the output as binary value which is 0 (yes) and 1(no). It finds the probability of [19]the relationship between one dependent variable and one independent variable using logistic function. Logistic Regression has 3 types and they are binary, multinomial and ordinal. LR is used in many daily applications like fraud detection, email spam detection and disease detection etc. The algorithm is easy to implement using the Linear Regression.

$$y = \beta 0 + \beta X1 + ... + \beta nXn \tag{2.3}$$

[6]The logistic curve is called sigmoid curve which is like S-shaped curve and it is used to find the sigmoid functional probability by

$$p = 1/1 + e^{-y} \tag{2.4}$$

where y = mx + c which does not give a linear curve. Sigmoid Probability on linear regression equation

$$p = 1/1 + e^{-\beta 0 + \beta X1 + ... + \beta nXn} \tag{2.5}$$
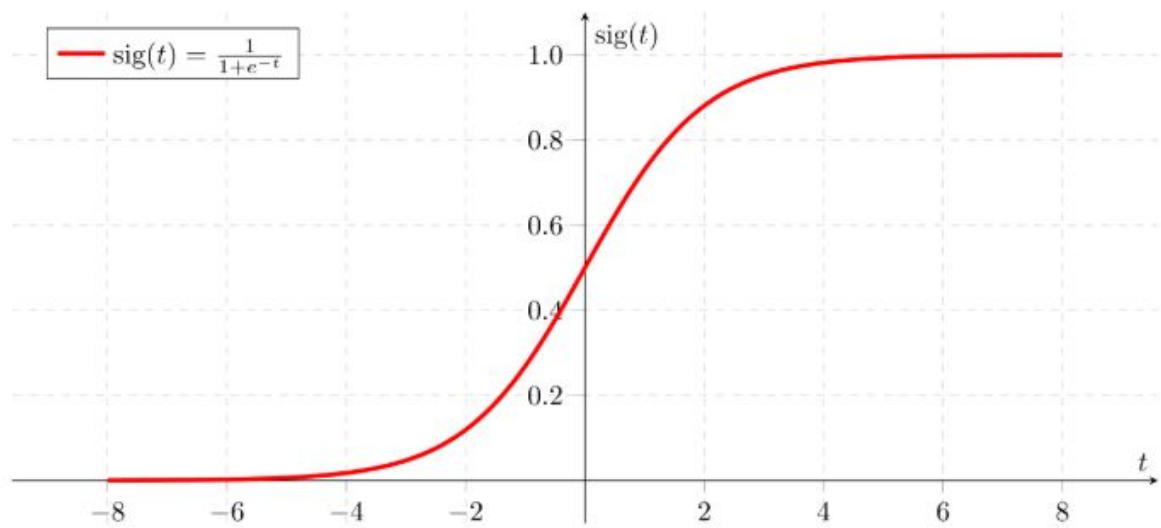
Figure 2.2: Sigmoid Curve

In figure 2.2 when the curve is in the positive axis it is predicted to be 1 and when the curve is in the negative axis it is 0. It has a threshold value from which it predicts to put 1 or 0 where it is appropriate i.e $0 \leq t \geq 1$. Therefore, the overall logistic regression is shown in figure 2.3.
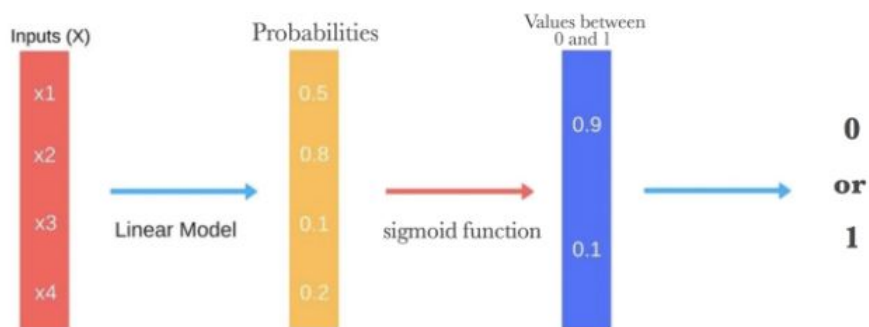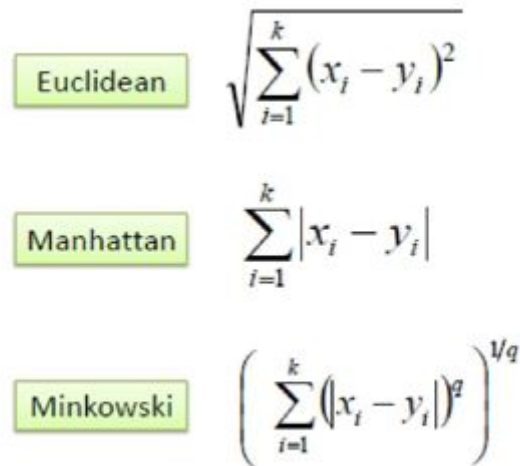


Figure 2.3: Logistic Regression

## K-Nearest Neighbor (KNN)

It is a supervised learning model and can be used for both classification and regression process. [8]KNN is a lazy learning and non-parametric algorithm because it does not need any training data rather it can take data directly from the dataset and use it in testing for making a model. Although it is slower and time consuming in testing, it is an easy and simple algorithm for finding the k closest value of the dataset. In KNN algorithm, first data from the dataset is taken to put it in training and testing where K value means the nearest or closest value will be selected. In order to select the k-closest value, the distance between points needs to be calculated. This can be done by different calculating methods such as Euclidean, Manhattan and Hamming distance where Euclidean distance calculating method is used frequently. After calculating the distance, arranging the data points in ascending order and then selecting the k nearest value and classifying the points. For reducing the error, repeat the process again and again and choose different k nearest values while maintaining the accurate predictions.

Euclidean $\quad \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\quad \sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\quad \left( \sum_{i=1}^{k}(|x_i - y_i|)^q \right)^{1/q}$
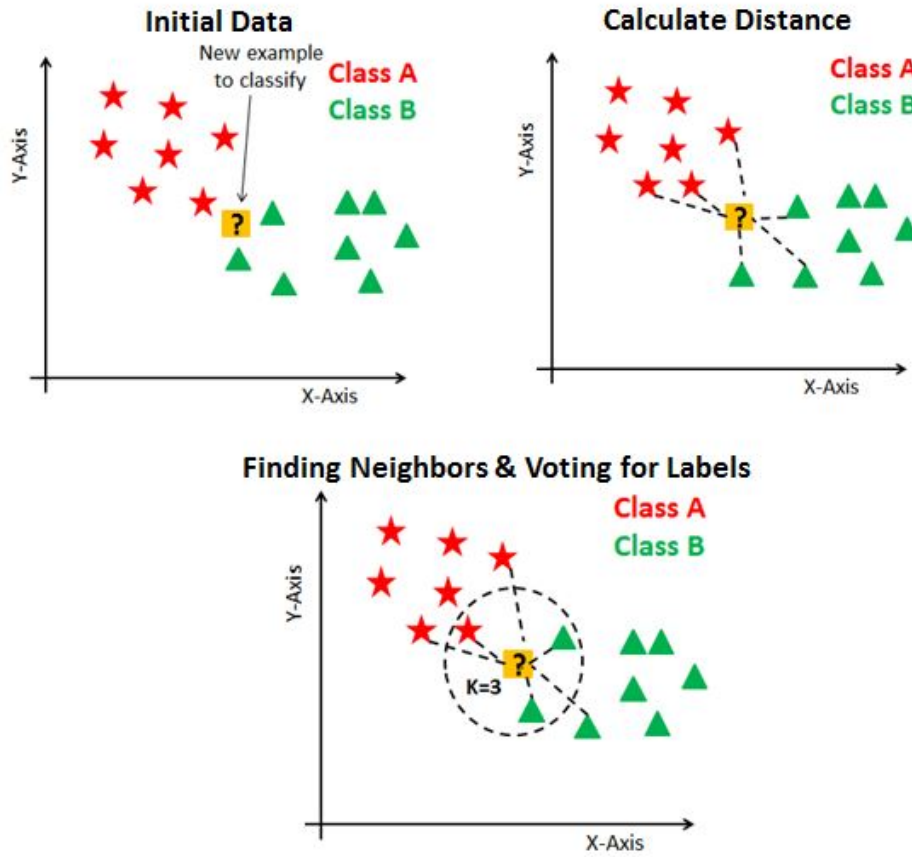
Figure 2.4: Formula for Calculating Distance

Figure 2.5: KNN Algorithm

**Support Vector Machine (SVM)**

SVM is a supervised machine-learning algorithm and it is a structural risk minimization which minimizes expected error of a learning model and for classification. It is highly preferable because it has many uses like face detection, image classification, text categorization and many more. [20] SVM uses kernel trick for non-linear Separation problems. SVM kernel has few types such as Linear, Polynomial and radial basis function kernel. [20] SVM is used to find a MMH and separates the data into classes.
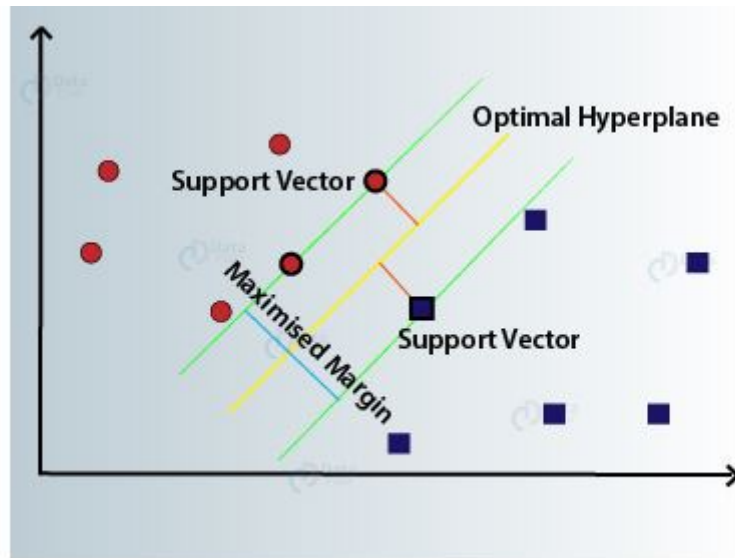
Figure 2.6: SVM

The above figure 2.6 shows how the SVM algorithm works in separating the two or more data by measuring the distance of the margin which means the two distant lines of nearest points and making a hyperplane line between the maximized margin lines of support vectors of the data points. This hyperplane is also called decision boundary. [20]The hyperplane can be made by segregating the classes in the best way and selecting the right hyperplane with the maximum segregation from the either closest data points as it is shown in the picture below in figure 2.7.
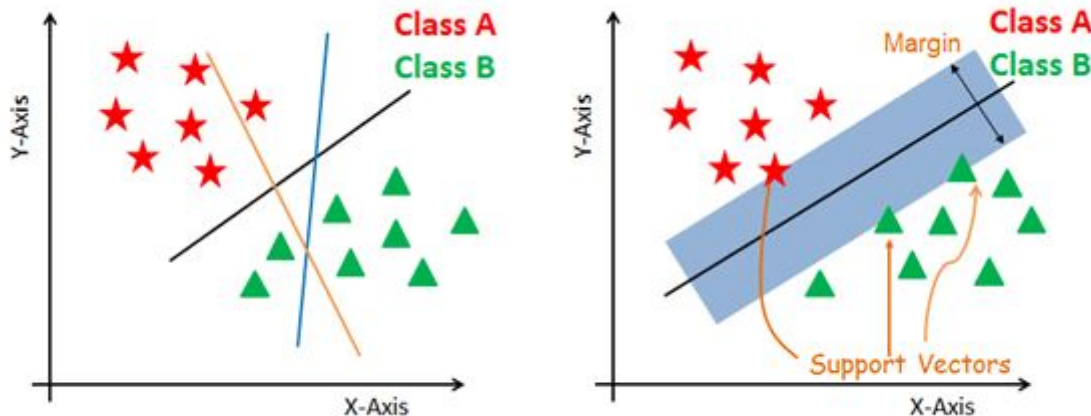


Figure 2.7: Creation of Hyperplane

14

## Random Forest

Random Forest is a supervised learning algorithm. It is based on the concept of Ensemble learning. [5]Mainly, it creates a tree or forest based on a decision tree method using data from the dataset and also predicts a result from each and every single tree and ultimately it draws a conclusion by selecting the best result after a voting process within the trees. Random forest mostly used in banking sector for the identification of loan risk, risks of the disease can be identified, identification of similar land use, stock market and many others. Random Forest uses the dataset to
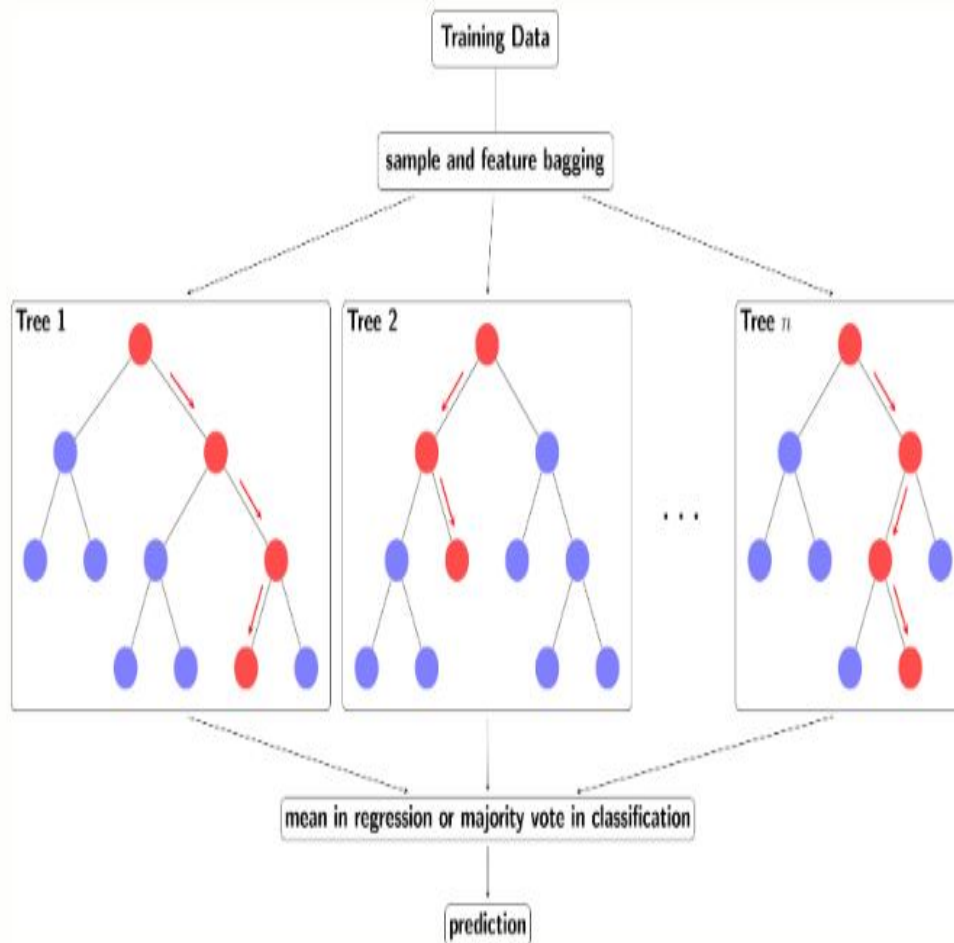


Figure 2.8: Random Forest

take the data by feature bagging which is randomly selecting data into different sets of the training data and from that data different decision trees will be generated and prediction will be on the majority of the vote. Repeatedly the method is running for better accuracy of the prediction. Therefore, the random forest method introduces more randomness and diversity. It also lets us determine how much the model is overfitting.

## 2.2.2 Algorithms for Recommendation

**Fuzzy KNN**

Fuzzy KNN is the improved version of the most famous supervised algorithm called KNN and this is because KNN is a simple, effective and easy classification method with no preprocessing requirement. Despite the useful facts of KNN, it has some flaws, that is if the class is unbalanced, it cannot give better performance. It means the data in the training set is not properly accurate and it may decrease the precision of classification. Hence, the theory of fKNN is constructed with improvement in order to have less error in classification. Fuzzy KNN algorithm [15]provides a more realistic vector of membership for the objects and accounts for the degree of object membership to the classes of objects. Fuzzy KNN is used in many applications such as developing a wine classification system, classifying web documents, color classification, a human action recognition problem, bankruptcy prediction, bioinformatics and in artificial immune recognition systems for treating disease diagnosis problems and many more. Basically, [15]class is assigned to the most common class. [15] fKNN algorithm is used mainly for dividing the cluster. .The theory of Warren and Damin are used to put membership into different categories.

$$u_i(x) = \frac{\sum_{j=1}^{K} u_{ij}\left(1/\parallel x - x_j \parallel^{2/(m-1)}\right)}{\sum_{j=1}^{K}\left(1/\parallel x - x_j \parallel^{2/(m-1)}\right)} \tag{2.6}$$

[13]Where i = 1, 2, …,c and j = 1, 2, …k, c is the number of classes, k is the number of k-nearest neighbours and m is the strength parameter. [13]Euclidean distance is between x and its jth closest points. uij is the degree of membership. The graph of Fuzzy KNN is shown in figure 2.9.
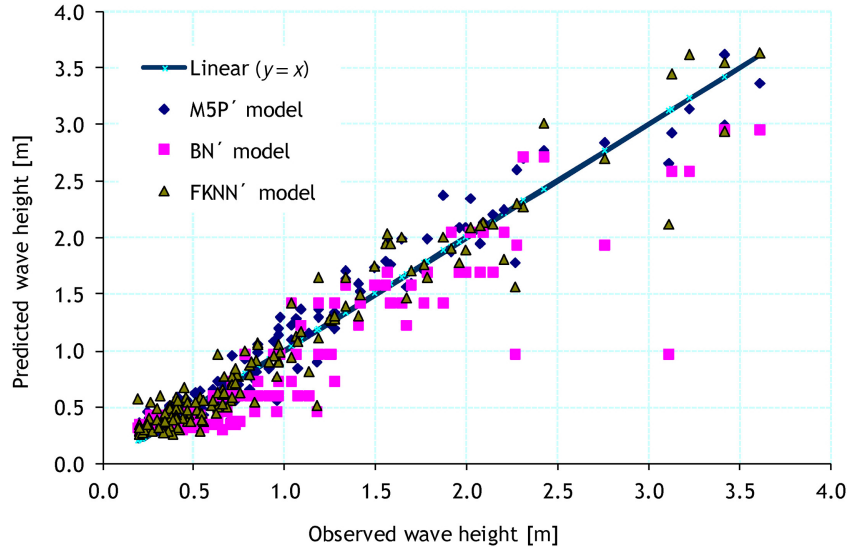


Figure 2.9: Fuzzy KNN

## Collaborative-Filtering Algorithm

Collaborative filtering is a popular recommendation algorithm that uses people's opinion to filtrate or evaluate the process. It is based on the idea that the system will predict a result on how many people have rated a similar item and it will be repeated or suggested again in future. For example if u think of a movie rating system then CF feedback falls into two options. Firstly, "Explicit" is when users specify how much they liked a particular item by providing a numerical rating. Secondly, "Implicit" when a user chooses a movie that he likes to watch for that moment, the system deduces that the user is interested in that particular kind of genre of movie. CF algorithm has many forms and they are user-based and item-based collaborative filtering. CF has many uses in our daily life and those are Amazon, Netflix, iTunes, IMDB and many recommendation or predicted systems. CF has many other types like memory based, model based. In our model, item based collaborative filtering is used for the purpose of agricultural productivity. It is because item based CF method filters items that are suitable for that particular parameters using the matrix factorization process and similarity calculation.
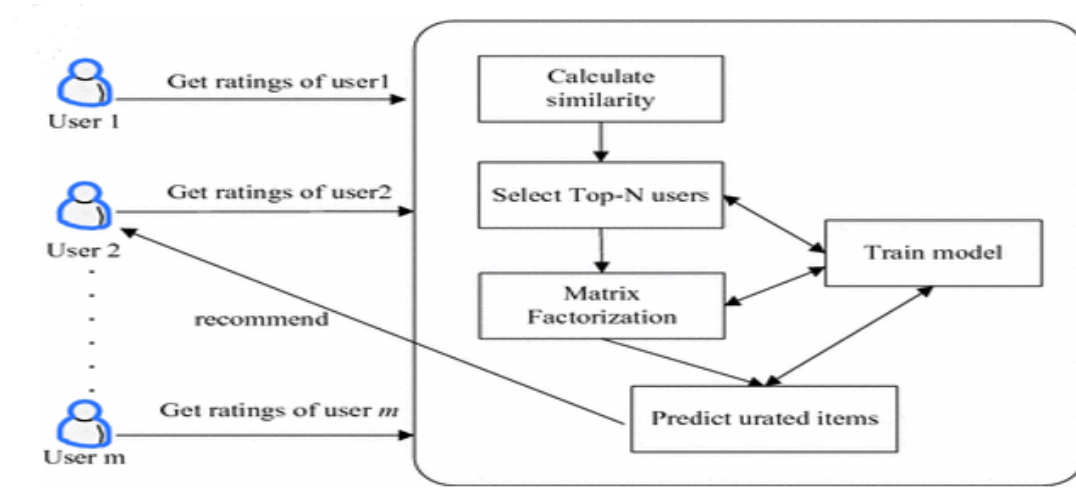


Figure 2.10: Collaborative Filtering

$$simil(x, y) = cos(\vec{x}.\vec{y}) = \frac{\vec{x}.\vec{y}}{\parallel \vec{x} \parallel \times \parallel \vec{y} \parallel} = \frac{\sum_{i \in I_{xy}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i \in I_x} r_{x,i}^2} \sqrt{\sum_{i \in I_y} r_{y,i}^2}} \qquad (2.7)$$

# Chapter 3

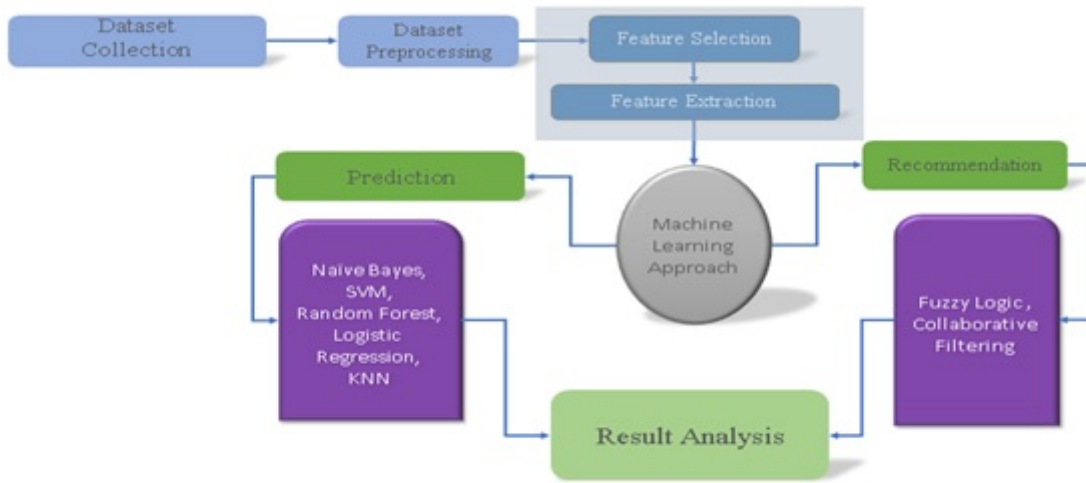# Proposed Model

## 3.1  Machine Learning Methodology



Figure 3.1: Proposed Model

The workflow of our approach is given in Figure 3.1 to predict the productivity of crops and recommend the best suitable crops according to the Climate feature like humidity, rainfall and temperature. We have taken many approaches to make our prediction and recommendation more effective, accurate and efficient. During this approach we have pre-processed our data, select the best feature, train our model and test them with test data for accuracy and then we predict and recommend best possible crops. After applying those prediction and recommendation we have used may charts for visualizing our data.

### 3.1.1  Dataset

Machine learning approach is a data-centric approach. Therefore, dataset plays a vital role in machine learning. We need to train our model to perform various actions by feeding data to our model. Our dataset is divided into two parts. One is training a dataset and another one is testing Dataset. Using training dataset, we train our

model with various algorithms and we validate our model using testing dataset and get the necessary result.

### 3.1.2 Dataset Description

We have taken our dataset from an Indian Agriculture website by email request. In this dataset we have about 7 years of data record which contained 135 of different crops based on various categories. In the dataset we have 12 columns and total 11690 rows. The column names are 'Statename', 'districtname', 'year', 'season', 'area', 'rainfall', 'humidity', 'temperature', 'prevYearsRainfall', 'prevYearsTemperature', 'prevYearsHumidity', and 'Crops'. For back-end engines we used Keras and TensorFlow. A heatmap of the dataset is given in the Figure 3.2:
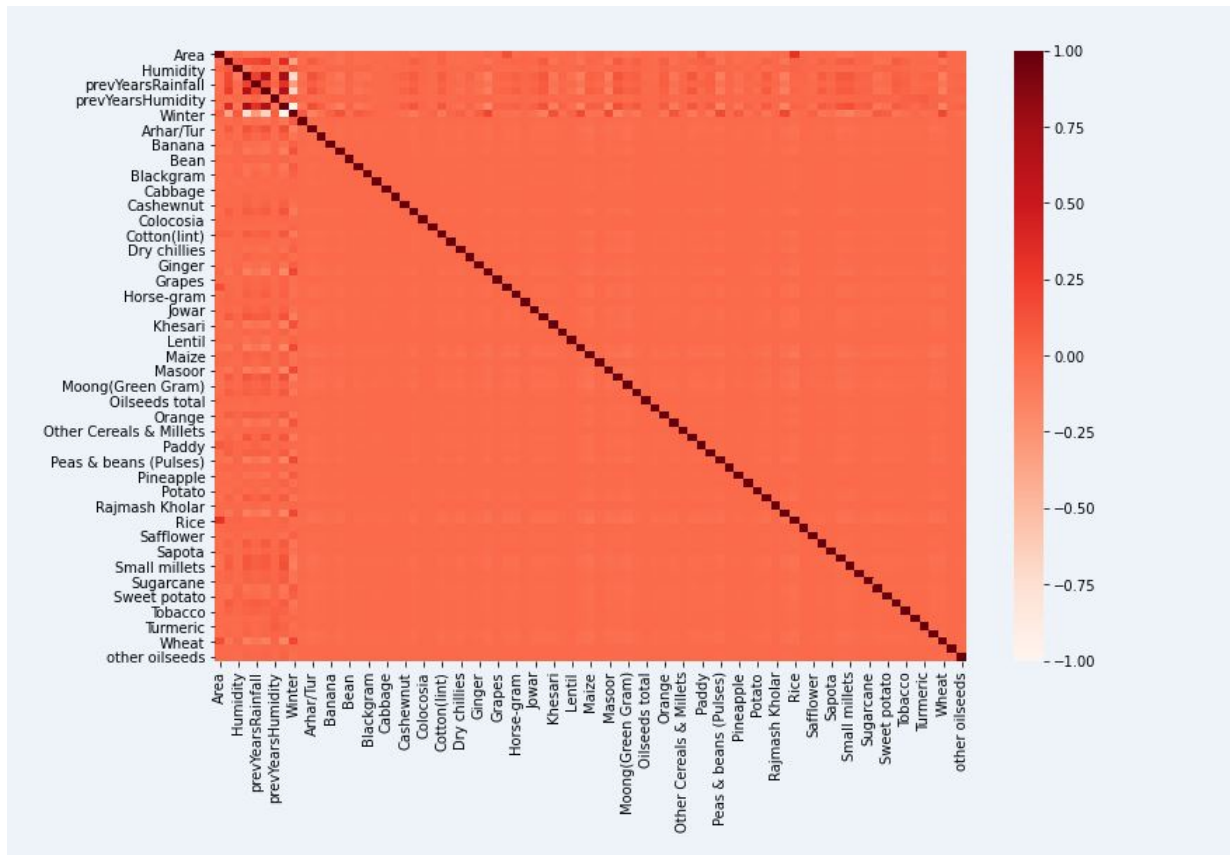


Figure 3.2: Heatmap

A distribution plot of the most often crops like 'Rice', 'Maize', 'Moong', 'Arhar', 'Urad', and 'Wheat' vs different parameters are shown in the Figure 3.3, Figure 3.4, Figure 3.5 :
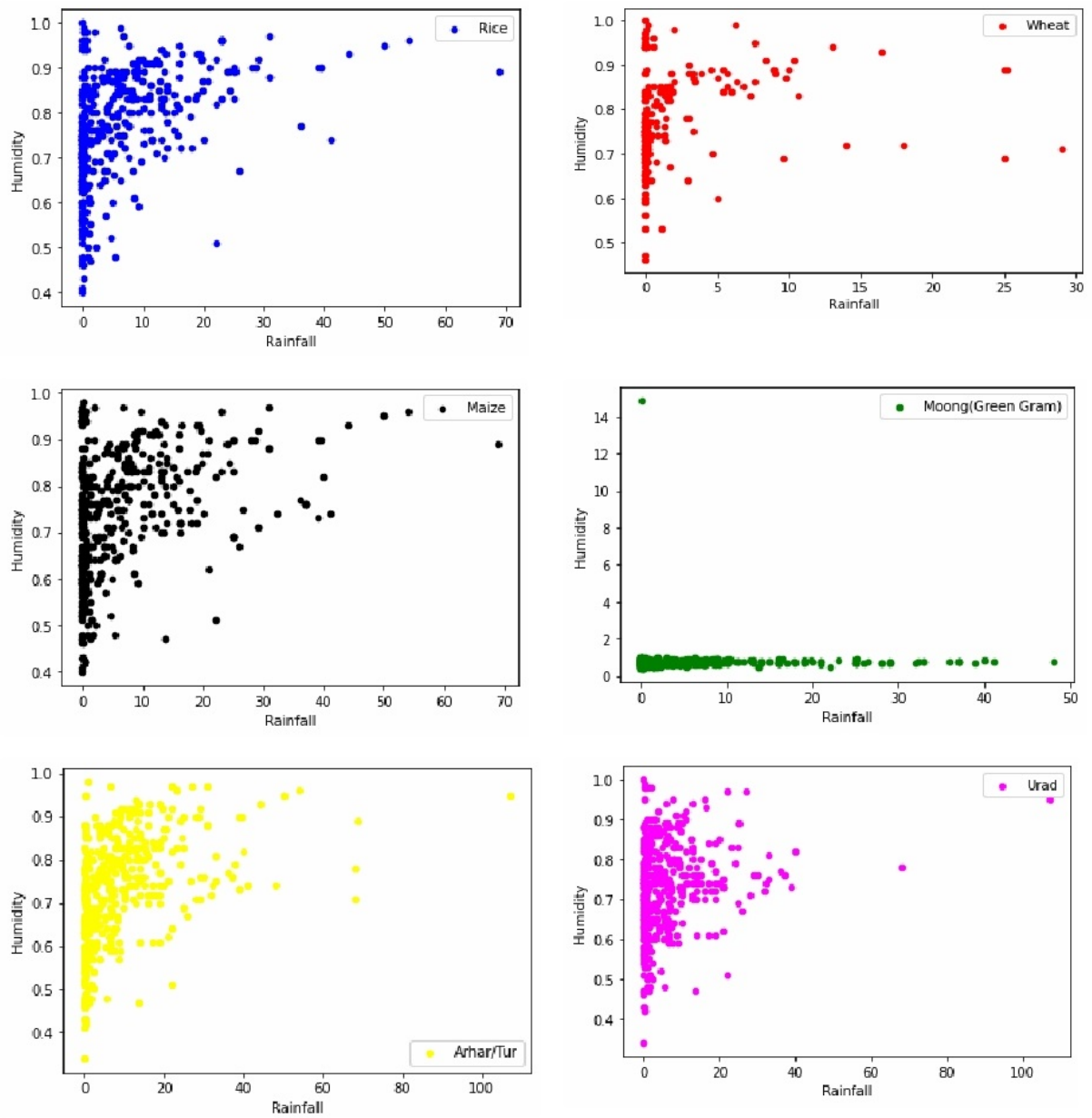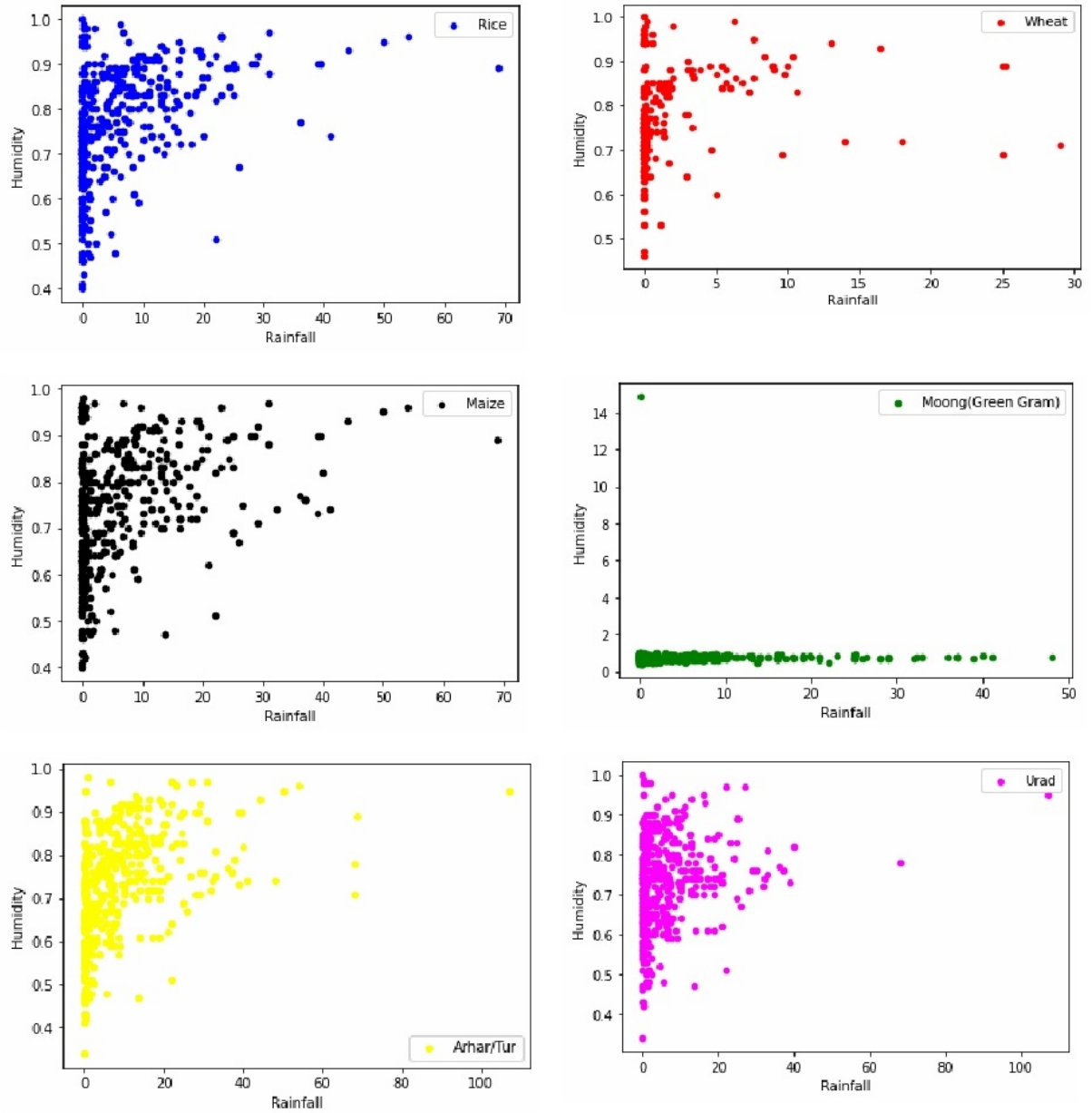
Figure 3.3: Crops vs Humidity and Rainfall

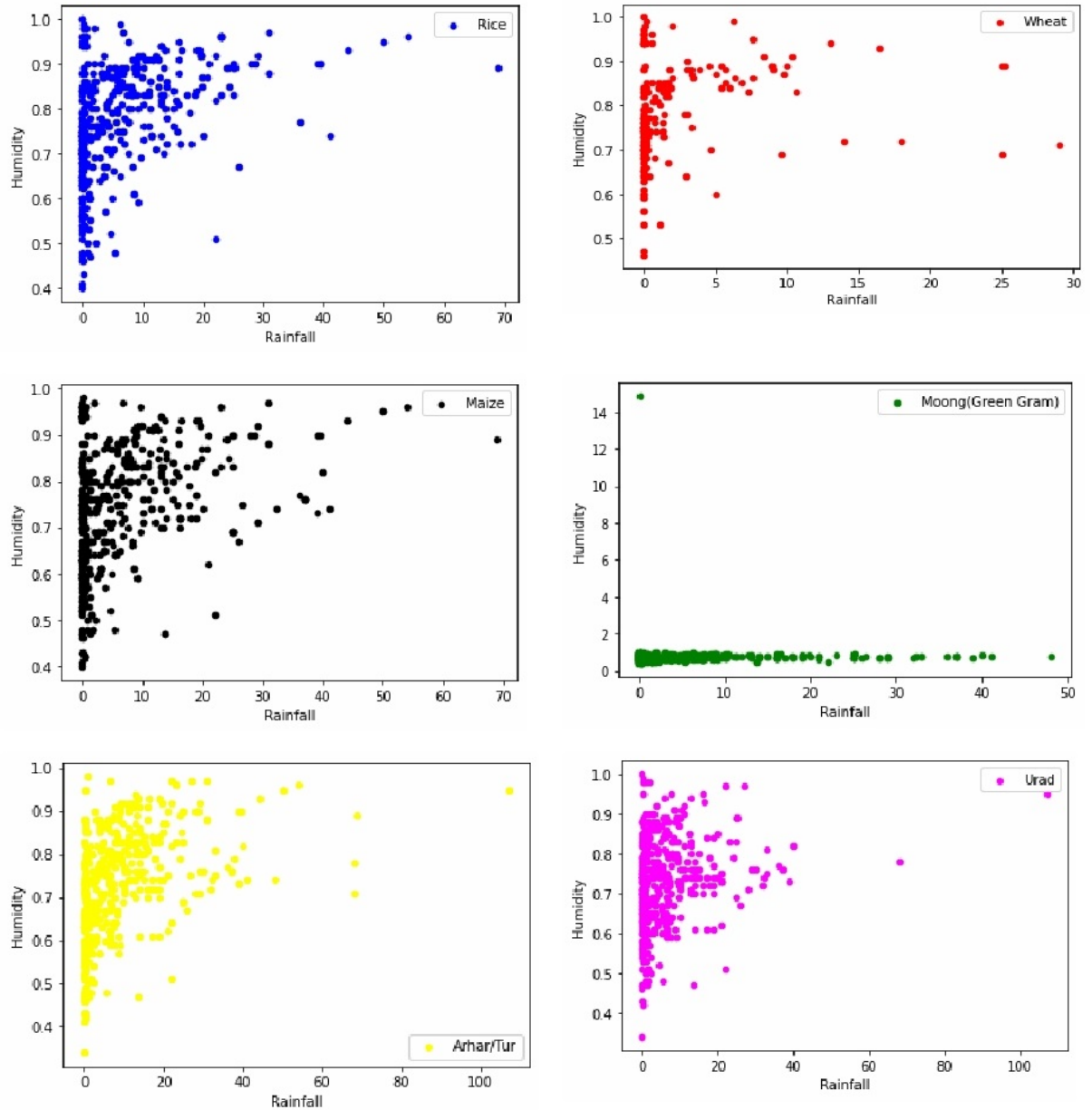Figure 3.4: Crops vs Humidity and Temperature

Figure 3.5: Crops vs Rainfall and Temperature

### 3.1.3 Data Pre-processing

Data Pre-processing is a process where any data is transferred to machine readable form. Machines are unable to interpret real world data. Therefore, we have to process our data such a way that it is understandable by machines.

Firstly, to process our data we need to import the necessary library. Libraries are the collection of functions which can be called and used when necessary. Using this library, we can import our dataset and apply many functions according to our necessity.

Secondly, we can import our dataset using the library. Generally, the dataset is in a csv form. We can read that csv data using panda's library. Using this library, we

can load our data into our machine..

Next, we need to reduce the unnecessary data from our dataset. Then, we have to check if there is any null value. If they exist then we can replace them by mean value. Again, there is lots of data which cannot be interpreted by machine. Therefore, if we want to work with them, we need to convert them into binary value.

Moreover, there are some categorical data which are tough for a machine to understand. Therefore, we need to encode the categorical data if we want to use that data into our model.

Now, we will divide our dataset into two parts, one is a training set and another one is testing set. We will build and try to train our machine learning models on our training set, i.e our machine learning models will try to find if there is any correlations between our training set and then we will test the models on our test set to check how accurately our model can predict the dependent variable. Generally, 80 percent data are used for testing and 20 percent data for training.

Now we need to divide our data into 2 sets, training and testing. Again, We will divide them into 4 sets—(1) X train (training part of independent variable), (2)X test (test part of the independent variable), (3) Y train (training part of the dependent variables based on the X train , Y test (test part of the dependent variables based on X test).We will split our data by 80 and 20 ratio or according to our algorithm choice

Finally, we need to scale our features. In Scaling we standardize our data ratio wise. In our model we need to use cosine similarity. Therefore, any negative value will give unreal value. We need to standardize the data to avoid negative value. To do this work, we will import the class Standard Scaler library from the scikit pre-processing library and use them when necessary. Now we will transform and fit our X train set. This will bring all the data into the same standard.
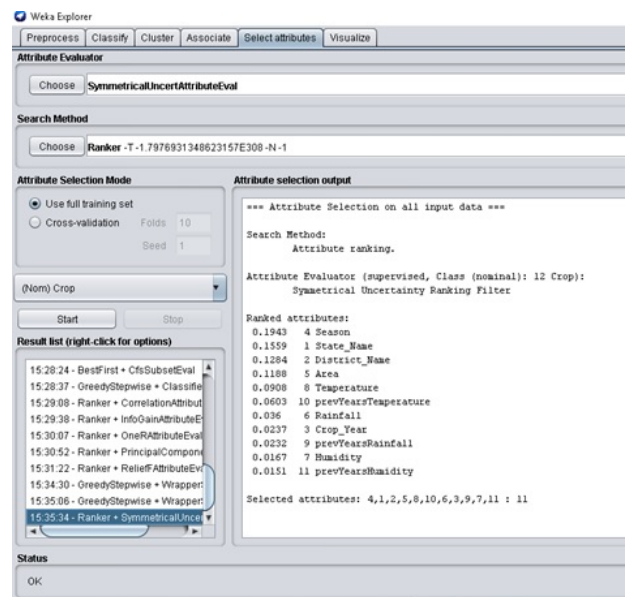
### 3.1.4  Feature Selection



Figure 3.6: Feature Selection using WeKa Explorer

Feature selection is the process of reducing the number of input variables when de-

veloping a predictive model. There are some model which have a lot of variable which can bring down the progress of our model. Again, data model which have an unnecessary variable can be a wastage of memory and speed. Moreover, Irrelevant input variable can degrade our model performance. In our dataset, State name, District Name, Crop year, Previous year humidity, Previous year rainfall, previous year temperature does not have any impact in our model. Again, we have also removed all the Null values using the feature selector.

For Selecting feature, we, have used a software called Weka Visuals. It is a software used for data mining. We can apply many algorithms in our data. It contains many tools so that we can pre-process our data. Weka is not only used for data preprocessing but also it can be used for classification, regression, visualization and data associative. We can use this software according to our need.

We have used "Symmetrical UncertAttribute Evaluation" which Evaluates the necessity of an attribute by measuring the symmetrical uncertainty with respect to the class. Ranker Ranks attributes by their individual evaluations. For selecting an attribute, we have used a full training dataset split into 80:20. However, we have used Season, Rainfall, Temperature, Humidity as a key attribute for our model.

### 3.1.5 Training and Testing of Machine Learning Classifier and Recommender

After we are done with our feature selection process, we will train and test our machine learning classifier. Then we will divide our dataset into 2 set which is training and testing sets. Next, we train and test the dataset using many classification algorithms. Training data are used to train and build the models. Test data are used for understand how well the model is performing. Test data is used to predict answers dependent on training. In our dataset, we used eighty percent data as training data and twenty percent as testing data in prediction and for the recommendation we used 50 percent data for training and 50 percent data for testing. After splitting the dataset, we train our model. With various algorithms which are given Below.
For Prediction:
1. naïve Bayes classifier.
2. K-nearest neighbor
3. Random Forest
4. SVM
5. Logistic regression.
This prediction algorithm tries to predict the crop based on our selected features
For Recommendation
1. Fuzzy Logic
2. Collaborative Filtering
These recommendation algorithms tries to recommend some best possible crops based on their selected features like the rainfall, humidity and temperature.

## 3.2   Proposed Prediction Model

When a user gives the different weather parameter values in the dataset it will store the data in the dataset and run these following 5 algorithms by splitting the whole
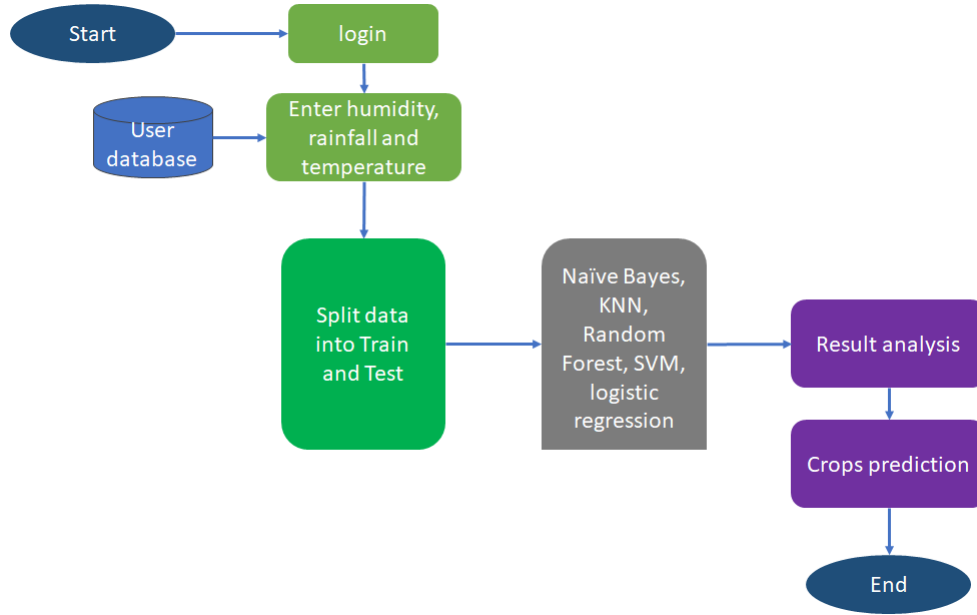
Figure 3.7: Proposed Prediction Model

data into 80% of the train and 20% of the test. After that, it will evaluate the result and show the best crops predicted.

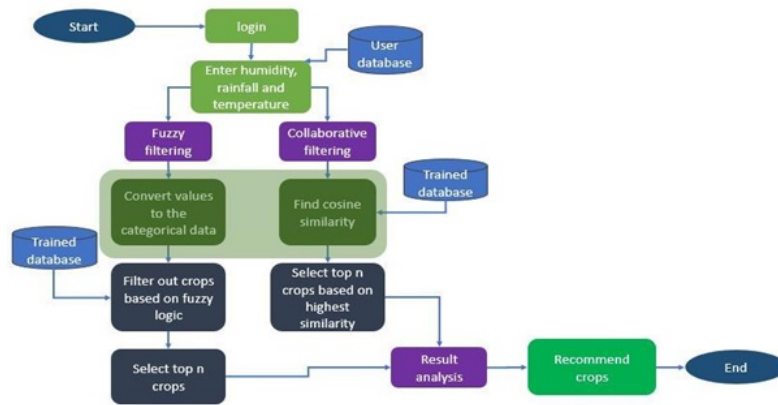## 3.3  Proposed Recommendation Model



Figure 3.8: Proposed Recommendation Model

In our recommendation model we have implemented two algorithms: Collaborative Filtering and Fuzzy K-nearest neighbor.

In Collaborative Filtering, Use will give their humidity, temperature, rainfall values. Then these values will be sent to our server. In Server we have our trained data. It will check the similarity between our trained data and input data. Then, it will filter out the 5 based crops based on their cosine similarity. Then, the input value with current crops will be updated in our model. Thus, it will improve the accuracy of our model day by day.

In Fuzzy K-nearest neighbor, User will give three inputs of humidity, rainfall and

temperature. Based on the value of this parameter the value will set under three categories 'High', 'Low', 'Moderate'. Then this data will be sent to our server. In our server, we have categorized our crops based on 27 conditions. When any of these 27-conditions match with our input, it will show the top 5 crops which are suitable based on a given condition. Our server will be updated continuously therefore; user will have a vast choice of crops based on their climatic condition.

# Chapter 4

# Result Analysis

## 4.1 Result Prediction on Algorithms

We implement five prediction algorithms and two recommendation algorithms, we also analyzed their precision. We got the highest precision in Naïve Bayse that

| Algorithms | Precision | recall | f1-score |
|---|---|---|---|
| Naïve Bayse | 72 | 60 | 58 |
| K-Nearest Neighbor(KNN) | 64 | 60 | 61 |
| Random Forest | 70 | 69 | 70 |
| Support Vector Machine (SVM) | 56 | 61 | 53 |
| Logistic Regression | 70 | 67 | 68 |

Table 4.1: Algorithm Precision

around 72. The lowest precision we got is 56 in SVM. For KNN we got precision of 64. Both Random forest and Logistic Regression we got precision of around 70. A comparison of the precision of these algorithms are given in figure 4.1.
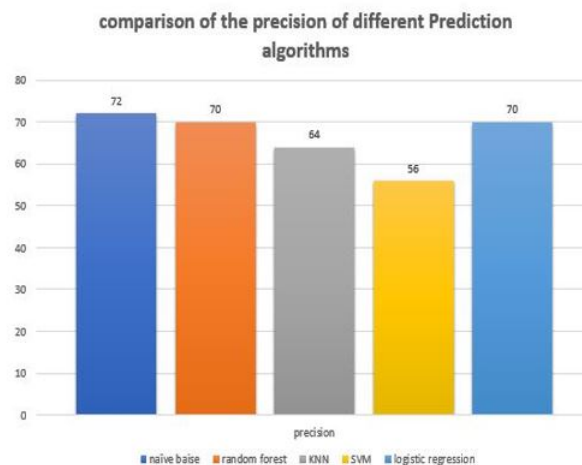


Figure 4.1: Precision comparison of different Prediction Algorithms

## 4.2  Predicted Crops



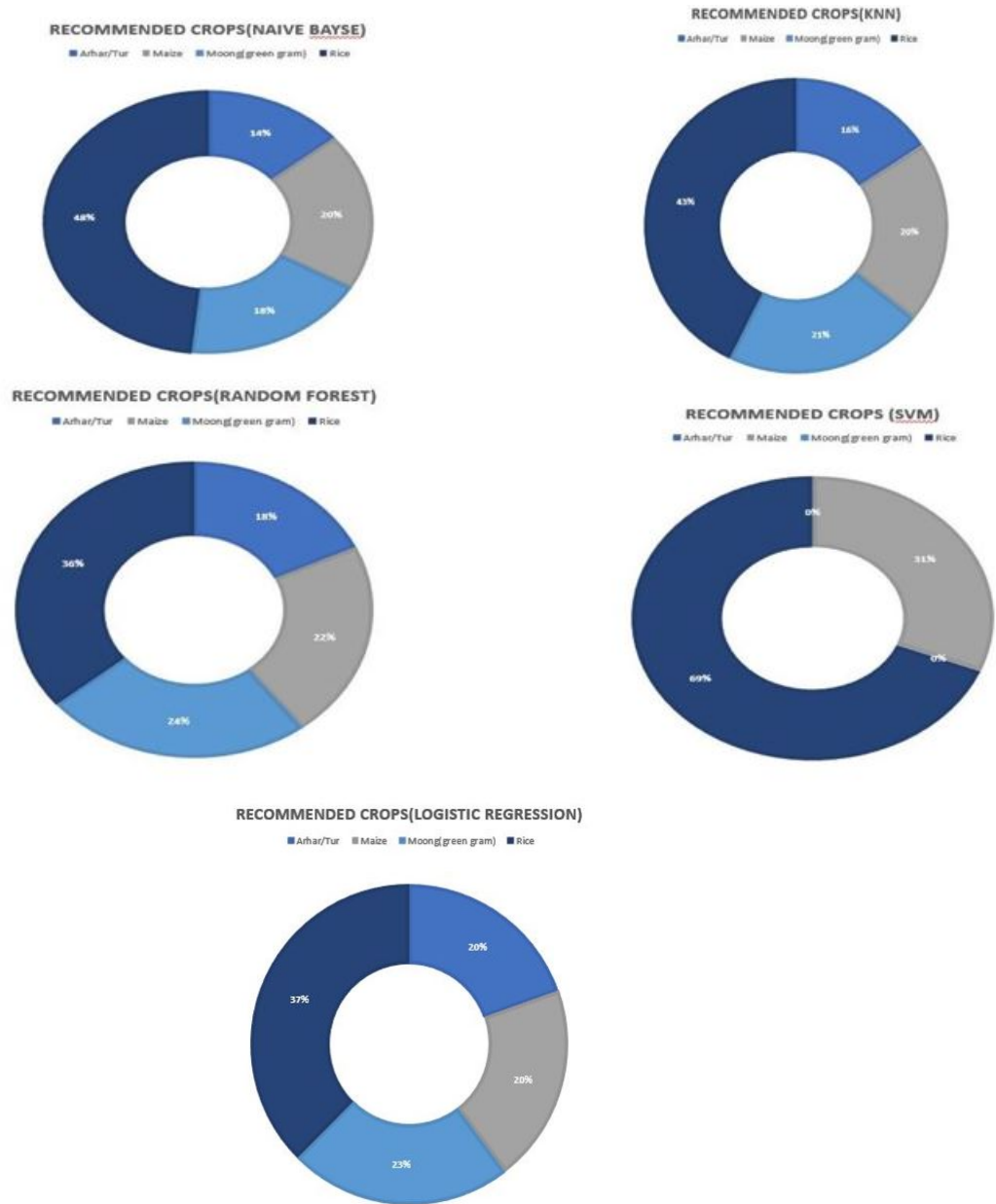Figure 4.2: Precision of the Top 4 Crops

We took the top 4 crops that have the highest precision in the following prediction algorithms and made a pie chat out of it. Here, Rice has the highest precision in every algorithm compared to other crops like Moong, Maize and Arhar/Tur. It is 48% in naive bayes', 43% in knn, 36% in random forest, 37% in logistic regression and greater portion of all in svm with 69%.

## 4.3 Collaborative filtering

Here, In our dataset in our dataset Rice were cultivated in 956 places, Bajra in 189 places, Urad in 472 places, Groundnut in 241 places, Maize in 900, Arhar in 478 places, Mung in 666 places. When we use cosine similarity to find similar geographical attribute then this crop had the best similar value. Using the similarity, we find that Maize has 2280 places with a similar attribute where it can be cultivated. For Bajra it has 445 places where it can be cultivated. Again, for Urad it can be cultivated in 1220 places. For Groundnut, it can be cultivated in 620 places. Moreover, Rice can be cultivated in 2450 places. Arhar can be cultivated in 1200 places. Lastly Mung can be cultivated in 1610 places according to our collaborative filtering.
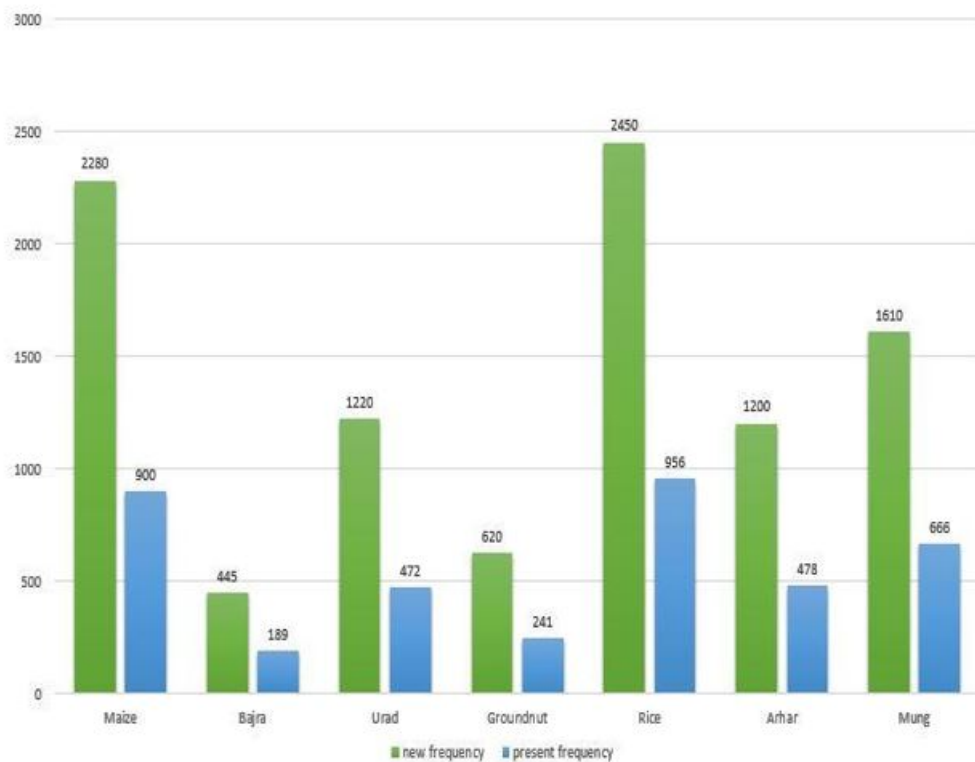


Figure 4.3: Bar Chart of co-sine fitting of the different kinds of crops in Collaborative filtering

## 4.4 Fuzzy filtering

We had a total $(3^3 = 27)$ combination because we have our 3 different input parameters: Rainfall, Humidity and Temperature and we divided them into 3 categories: Low (L), Moderate (M) and High (H). While filtering our dataset we got output on only 5 combinations:

| Fuzzy combinations | Recommended crops |
|---|---|
| Rainfall = low<br>Humidity = low<br>Temperature = low | Maize, Rice, Wheat |
| Rainfall = low<br>Humidity = low<br>Temperature = moderate | Rice, Wheat, Gram |
| Rainfall = low<br>Humidity = low<br>Temperature = high | Maize, Rice, Arhar/Tur |
| Rainfall = low<br>Humidity = moderate<br>Temperature = moderate | Sunflower |
| Rainfall = moderate<br>Humidity = low<br>Temperature = moderate | Rice, Jute, Maize |

Table 4.2: Fuzzy Filtration

# Chapter 5

# Conclusion and Future Work

In this chapter future works and a brief review of current work has been discussed. In future we have a hope that we design a web framework through Python Flask where it is easy to put input and see the desired result. We plan to use Python Flask because it can be accessed by the user and it helps the user to find the desired outcome and their expectations.

We also have a hope to increase our dataset, for now we have only worked on external factors such as rainfall, humidity, temperature. In future we want to include more factors in our dataset such as the fertility of the land. We want to predict possible diseases or the probable attacks of pests that can affect the production with sufficient parameters. We also add more crops to our recommendation as it will suggest the cultivators to go for much better alternatives.

To conclude, for our problems we have applied the algorithm for prediction and recommendation. For prediction we have applied the algorithm Naïve Bayes Classifier (48% for Rice), Support Vector Machine (69% for Rice), K-Nearest Neighbor (43% for Rice), Logistic Regression (37% for Rice), Random forest (36% for Rice). In this way we have predicted the productivity where the crop rice shows the maximum in every algorithm. And for recommendation using collaborative filtering it shows the recommended crops Rice, Maize, Moong as three most effective crops. On the other hand, fuzzy logic filters the crops according to the condition of the crop yielding factors. In this way we have worked on our ideas to develop. Our system solves the problems with limited parameters. So, increasing the parameters and adding more attributes to our system will be credible for the problems which were mentioned before. In this way we hope for the advancement of our agricultural economy, increasing productivity and contribution to the prosperity of the country.

# Bibliography

[1]  S. S. Dahikar and S. V. Rode, "Agricultural crop yield prediction using artificial neural network approach", *International Journal Of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, vol. vol. 2, no. 2, Jan.2014.

[2]  V. Lamba and V. S. Dhaka, "Wheat yield prediction using artificial neural network and crop prediction techniques (a survey)", *International Journal For Research In Applied Science and Engineering Technology (IJRASET)*, vol. vol. 2, no. 9, Sep.2014.

[3]  M. S. D. P. Kaviani, "Short survey on naive bayes algorithm", *International Journal of Advance Engineering and Research Development*, vol. 4, 11 Nov.2017.

[4]  *Bangladesh - agricultural land (% of land area)*, [Online]. Available: https://tradingeconomics.com/bangladesh/agricultural-land-percent-of-land-area-wb-data.html, (Accessed: 2016).

[5]  A. Navlani, *Understanding random forests classifiers in python*, [Online]. Available: https://www.datacamp.com/community/tutorials/random-forests-classifier-python, (Accessed: May 16, 2018).

[6]  N. S. Chauhan, *Real world implementation of logistic regression*, [Online]. Available: https://towardsdatascience.com/real-world-implementation-of-logistic-regression-5136cefb8125, (Accessed: March 11, 2019).

[7]  R. Van Meteren and M. Van Someren, "Using content-based filtering for recommendation", in *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*, vol. 30, 2000, pp. 47–56.

[8]  A. Navlani, *Knn classification using scikit-learn*, [Online]. Available: https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn, (Accessed: August 02, 2019).

[9]  M. Begum and L. D'Haese, "Supply and demand situations for major crops and food items in bangladesh", *Journal of the Bangladesh Agricultural university*, vol. 8, no. 1, pp. 91–102, 2010.

[10]  T. Sujjaviriyasup and K. Pitiruek, "Agricultural product fore-casting using machine learning approach", *Int. Journal of Math. Analysis*, vol. 7(38), pp. 1869–1875, 2013.

[11]  A. A. Chirde and U. K. Biradar, "A survey on collaborative filtering in accordance with the agricultural application", *International Journal of Computer Applications*, vol. 975, p. 8887, 2014.

[12] D. Zingade, O. Buchade, N. Mehta, S. Ghodekar, and C. Mehta, "Crop prediction system using machine learning", *Int. J. Adv. Eng. Res. Dev. Spec. Issue Recent Trends Data Eng*, vol. 4, no. 5, pp. 1–6, 2017.

[13] Z. Cai, J. Gu, C. Wen, D. Zhao, C. Huang, H. Huang, C. Tong, J. Li, and H. Chen, "An intelligent parkinson's disease diagnostic system based on a chaotic bacterial foraging optimization enhanced fuzzy knn approach", *Computational and mathematical methods in medicine*, vol. 2018, 2018.

[14] B. K. R. M. Kuanr and S. N. Mohanty, "Crop recommender system for the farmers using mamdani fuzzy inference model", *International Journal of Engineering Technology*, vol. 7, 2018.

[15] M. R. A. Md. R. Nikooa R. Kerachian, "A fuzzy knn-based model for significant wave height prediction in large lakes", *Oceanologia*, vol. 60, 2 2018.

[16] N. Kim, K.-J. Ha, N.-W. Park, J. Cho, S. Hong, and Y.-W. Lee, "A comparison between major artificial intelligence models for crop yield prediction: Case study of the midwestern united states, 2006–2015", *ISPRS International Journal of Geo-Information*, vol. 8, no. 5, p. 240, 2019.

[17] K. Moharm, "A framework for adaptive personalized e-learning recommender systems", *International Journal of Intelligent Information Systems*, vol. 8, no. 1, pp. 12–17, 2019. DOI: 10.11648/j.ijiis.20190801.13.

[18] *Bangladesh: Growing the economy through advances in agriculture*, [Online]. Available: https://www.worldbank.org/en/results/2016/10/07/bangladesh-growing-economy-through-advances-in-agriculture, (Accessed: October 9, 2016).

[19] N. Donges, *The logistic regression algorithm*, [Online]. Available: https://www.experfy.com/blog/the-logistic-regression-algorithm, (Accessed: January 08, 2019).

[20] A. Navlani, *Support vector machines with scikit-learn*, [Online]. Available: https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python, (Accessed: December 27, 2019).

[21] J. Vincent, *Chinese farmers are using ai to help rear the world's biggest pig population*, [Online]. Available: https://www.theverge.com/2018/2/16/17019446/farming-ai-pig-tracking-china-alibaba, (Accessed: February 16, 2018).

[22] *Bangladesh: Rising temperature affects living standards of 134 million people*, [Online]. Available: https://www.worldbank.org/en/news/press-release/2018/09/26/bangladesh-rising-temperature-affects-living-standards-of-134-million-people, (Accessed: September 26, 2018).