Course Project_Final Report

Subject: Data and Web Mining (CSE 420)

Topic: Crop Recommendation Prediction System

➔The crew of the Project:

- AP19110010174 - Boyapati Sai Venkat.
- AP19110010192 - Konakanchi Subrahmanyam.
- AP19110010168 - Maddirala Sai Karthik.
- AP19110010216 - Suryadevara Sai Bhuvanesh
- AP19110010199 - Somepalli Sagar Suji

Topic:

Crop Recommendation Prediction System

Abstract

Economic growth and employment in India are heavily dependent on agriculture. Globally, agricultural research has strengthened the optimization of profits and is an important and very vast domain to gain more advantages. Today, there are countless people with the land, but they are not aware of how to yield crops. Agriculture is the only way for everyone to prosper in the future. Most people cultivate crops on improper soil and, as a result, are doing useless agriculture.

Farmers in India commonly choose crops that are not suited to their soils, which causes a setback in productivity. Precision agriculture has addressed this problem of farmers by offering the right crops according to the soil conditions. There are many ways to give recommendations in India, which is considered to be an agricultural country. Today's recommendations are based on farmer-to-farmer communications, with expertise coming from a variety of sources.

Precision farming is a modern farming technique that is based on the collection and analysis of data on soil characteristics, soil types, and crop yield and suggests the correct crop to farmers according to site-specific parameters. By doing so, the wrong crop is less likely to be chosen and productivity is increased. It is believed that this structure could predict the best harvest for the agronomist's area. Furthermore, suggestions are made for various farming practices and strategies including diversity of crops, spacing, irrigation, sow processing, etc., as well as fertilizer and pesticide suggestions. These are based on historical soil indices

and estimates of crop and weather costs.

In this application, you will be able to identify the types of soil, the water source of that land, and what crop you should plant. You will also be able to suggest what crops will grow well on that soil. So the application provides information about agricultural people. For instance, we can determine the crop which is most suitable for the soil, weather conditions, temperature, etc. Therefore, the crop for each soil is determined based on machine learning. Using past agricultural activities data, recommendations can be provided to farmers on what fertilizer to use and what crop they should plant. This application can also be used to increase crop yields and recommend that you plant certain crops based on your past agricultural activities.

Key Words

Crop Recommendation Prediction system, R2, Data Preprocessing, Dataset, KNN, Linear Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, Root Means Square Error, Mean Square Error, Mean Absolute Error.

Introduction

The area under agriculture is more than 1.6 million square kilometers, which makes it the second-largest country in the world. Agriculture practices are still being practiced in India, one of the oldest countries in the world, though they have drastically evolved in recent years as a result of globalization. The vast majority of Indians are engaged in agriculture and so are dependent on the economy of agriculture. India is poised to become an agricultural superpower. As much as 55% of the Indian population depends on farming, while in the US, a small percentage of the total population is dependent on farming. Rain-fed agriculture is the most common method of growing food in India, with just about 45% of the land irrigated.

Indian agriculture has been adversely affected by a variety of factors. India's agricultural sector suffers from a lack of attention, leading to a decline in the number of farmers living on farms and leading to a rise in suicides among farmers. There is no such universal aid system for farmers in agricultural India. To make more money, farmers must increase productivity so they can work less and get more from the same piece of land. In order to regain agriculture's health, many new technologies have been developed. So we can make recommendations using the rich collection of agricultural data we have collected previously. Aids in reducing poverty and promoting rural development through agriculture.

Today, farming professionals are taking part in growing crops in demand on the market that is currently in demand. In blindly following such an approach, however, they neglect to consider issues like their land is incompatible with such crops or it lacks the resources that the crop needs, resulting in less yield. A precision farm can help in this situation. There is a huge amount of research being conducted to develop a more accurate and efficient crop forecasting model. This research field includes various machine learning techniques like assembling. Among these are various machine learning techniques such as enchantment.

A process known as precision farming focuses on applying precise and appropriate amounts of substances like fertilizers, soil, etc. but in recent times, the way in which agriculture is done has drastically changed due to globalization. The advantage of precision agriculture is that we've been able to achieve efficiency in input and output and make better decisions regarding agriculture. A major domain of precision agriculture is the recommending of crops. This is dependent on many factors. Systems can recommend various crops, fertilizers, and farming techniques. It is the aim of precision agriculture to identify these parameters in a site-specific manner which will allow issues concerning crop selection to be resolved.

Agriculture in India has been affected by several factors. Many new technologies are being developed to restore health. There are many types of precision agriculture, including those which increase yields and productivity by applying the right amount of pressure at the right time to the crop. Not all precision agriculture systems are effective. Precision farming is an approach that is site-specific.

Literature Reviews

A major issue for Indian farmers is not choosing the proper crop for their land. Agriculture is India's primary source of revenue and employment. Due to this, they will have to reduce their production significantly. Farmers have benefited from precision agriculture by overcoming their obstacles. It is a method of crop production that seeks to inform farmers of the optimal crop for their specific site using research data on soil characteristics, soil types, and crop production statistics. Consequently, fewer crops are selected incorrectly and production is increased. [1]

Crop yield forecasts are important for farmers, government agencies, and researchers to use for making informed decisions about crop storage, selling, setting minimum support prices, and importing and exporting.

This prediction method is suitable for data mining since crop prediction requires an extensive study of a variety of factors such as soil quality and pH, EC, N, P, and K. This method of prediction is optimized for data mining since a great deal of data is needed. A data mining technique is used to extract knowledge from large quantities of data. This report discusses several methods for estimating agricultural yields using data mining. The success of any crop production prediction system depends on the accuracy with which traits are extracted and the effectiveness of the classifier. An overview of agricultural output forecast algorithms, their accuracy, and suggestions are provided in this paper. [2]

Agricultural decision support systems can offer scientific foundations for agricultural research as well as guidance for farm production. Intelligent decision support systems are benefitted from implementing big data analytics.

Researchers and developers are exploring the development of agricultural intelligent decision systems. Agricultural decision systems are categorized for the first time, the frame designation and design method for intelligent decision systems are examined, and the category is presented for the first time. [3]

Agricultural production is extremely important to India. Farmers prosper and so does the country. As a result of our efforts, farmers are able to grow the appropriate seed based on soil conditions, improving the country's output. By adding more characteristics and incorporating yield predictions, we intend to improve the data set further. [4]

Precision farming is described in detail, including the requirements and planning required for its development. The paper presents an overview of the basics of precision farming. Beginning with the basics of precision farming and moving towards developing a model that can support it, the author develops a framework.

To improve the rate of variability control over small, open farms, this paper illustrates the application of Precision Agriculture (PA) principles. In addition to providing direct advisory services via SMS and email, the model aims to reach even the tiniest farmer at the level of his/her smallest plot of the crop. Hence, this model can be implemented elsewhere in India with minor changes, only because it was built for Kerala State where the average holding size is much smaller than most of India. [5]

Data mining and visual data mining techniques are used to analyze agricultural data to acquire useful knowledge about yield, and fertilizer application. This paper reduces the high dimensional agricultural data to a more manageable size to help obtain useful information about yield, and fertilizer application.

The results show that Self-organizing maps are most suitable for large datasets and multi-dimensional scaling is most suitable for small datasets. Self-organizing maps are used to reduce the data, as well as multi-dimensional source scaling techniques to reduce the data. [6]

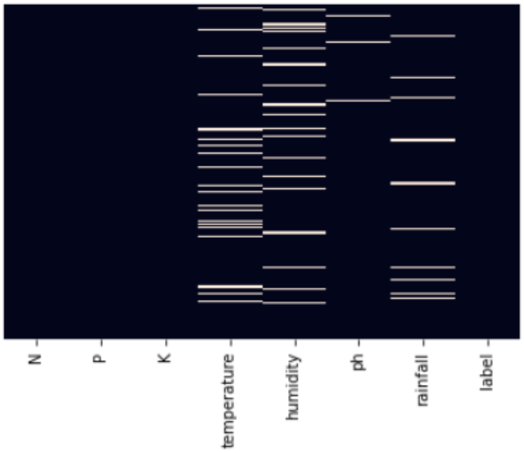
Data Pre Processing and Analysis

Data pre-processing is a critical step in transforming a large-scale dataset into a useable format. To narrow down the training of the model to the right level, this encompasses eliminating the NaN values (Missing Values) and inappropriate noisy data in the dataset.

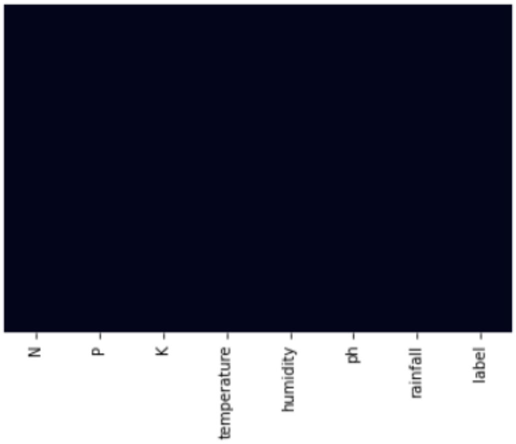
To make our dataset acceptable for training the ensemble learning model, we removed records that contained NaN values.

For Crop Recommendation Dataset

I. Before

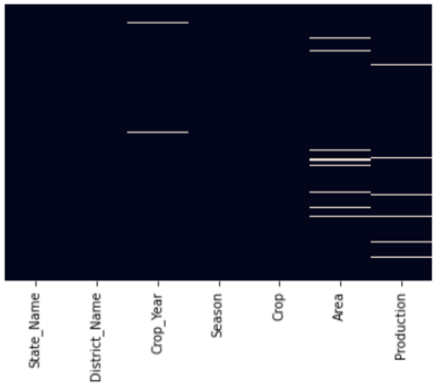


II. After

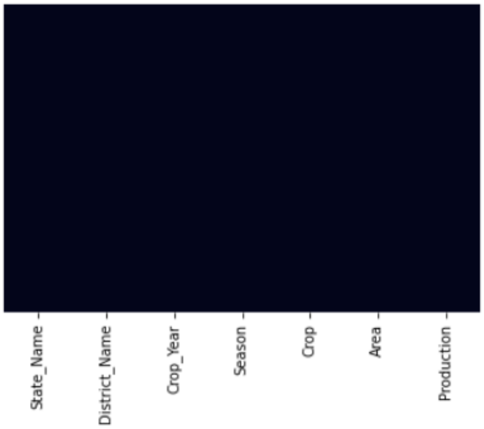


For Crop Yield Prediction Dataset

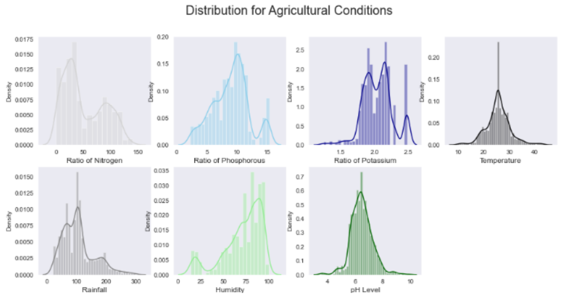
I. Before



II. After



Distribution of Agriculture



Proposed Methodology

Linear Regression

In general, a linear regression model is a statistical analysis method that uses regression analysis in mathematics to determine the quantitative relationship between two or more variables. It represents the relationship between one or more independent variables and dependent variables using a least-squares function called a linear regression equation. A linear combination of one or more regression coefficients, which are model parameters, makes up this function. When there is just one independent variable, simple regression is employed, however, multiple regression is used when there are several independent variables. Linear regression has a variety of applications.

Linear regression is a model that predicts the proportionate relationship between a dependent variable and a predictor by fitting the mapping between data input and output. The traditional linear regression equation is as follows:

$$y = \alpha + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn + \varepsilon$$

The beta represents the estimated parameter for the independent variables xi, the y represents the dependent variables, and the if represents the computation error term. While estimating the parameters the purpose of the linear regression model is to minimize the sum of squared errors.

Decision Tree

The decision tree is a well-known machine learning technique based on the principle that the same (similar) input creates the same (similar) output. The goal of making decisions based on tree outcomes is to classify or regress samples with the same characteristics by assessing the judgments of the samples' different attributes and categorizing them into the next leaf node.

The decision tree is a method of categorizing data using a set of criteria. It uses a rule-based technique to determine which values will be received under certain circumstances. For discrete data, there are classification trees, while for continuous variables, there are regression trees.

KNN

K-Nearest Neighbor (KNN) is a regression and classifying predictive analysis machine learning method. It's also known as a lazy learner algorithm because it doesn't study the data it's trained on, instead, it classified the new data it's tested on based on its similarities. KNN uses feature similarity to estimate house price values, which assigns a value to new data based on how closely it relates to the points in the training set.

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

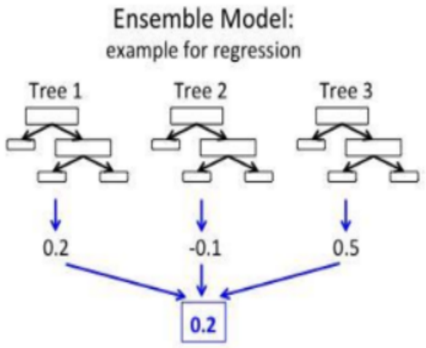
Where X is the new argument, Y denotes the existing point, and K indicates the K-Factor (number of clusters evaluated by the algorithm before allocating a value).

It's a supervised learning algorithm that's easy to use, versatile, and implement. It is based on the idea that creating a bond is near in proximity. To express the concept of similarity, it calculates the distance between two points on a graph. The numerical parameter 'k' in the algorithm shows how many data points should be taken into account when voting. We encircle a new point with K number of data points and assign it to the group with the most points within the circle to classify it. The best technique to figure out the value of K is to try out a few different values before settling on one, which decreases the error while maintaining the prediction's accuracy.

Random Forest

The random forests method is a decision tree ensemble method that is hard to learn and overfits easily. It employs a method known as tree bagging. By designing the decision tree, the problem of decision tree instability and excessive variety can be addressed. Because they are built using random sampling methods, these decision trees are referred to as random forests.

The method works by randomly selecting a subset of explanatory factors and training weak learners on their own. It creates a prediction model by simultaneously training weak prediction models, such as decision trees. A model averaging strategy is used to integrate the forecasts of each tree. The random forests method works by randomly selecting a set of characteristics and constructing a decision tree on its own. The decision trees of random forests are independent of one another. Majority voting can be used on these types of trees.



Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised learning technology with a lot of power. To find a hyperplane for data segmentation, the Support Vector Machine (SVM) technique turns the original data into a high-dimensional space. Support vectors, also known as "essential training tuples," define the hyperplane. SVM is more accurate than other algorithms because of its ability to accept nonlinear boundaries.

SVM models are implemented using two sets of input variables. The first is based on Systematic subset selection's five fundamental components. The second input set consists of six PCA transformation components.

Naive Bayes

It operates by applying the Bayes theorem to data and making the naïve assumption of conditional independence between every pair of characteristics, given the class variable's value.

Root Mean Square Error (RMSE)

The model developed in this study will be tested using the Root Mean Square Error (RMSE). The RMSE is used to compute expected performance by taking into account each data's prediction error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - p_i)^2}$$

Mean Square Error

The most basic and regularly used loss function is the Mean Squared Error (MSE), which is frequently taught in Machine Learning courses. To calculate the MSE, divide the difference between your model's predictions and the ground truth, square it, and average it over the entire dataset. The MSE will never be negative because we are constantly correcting our errors.

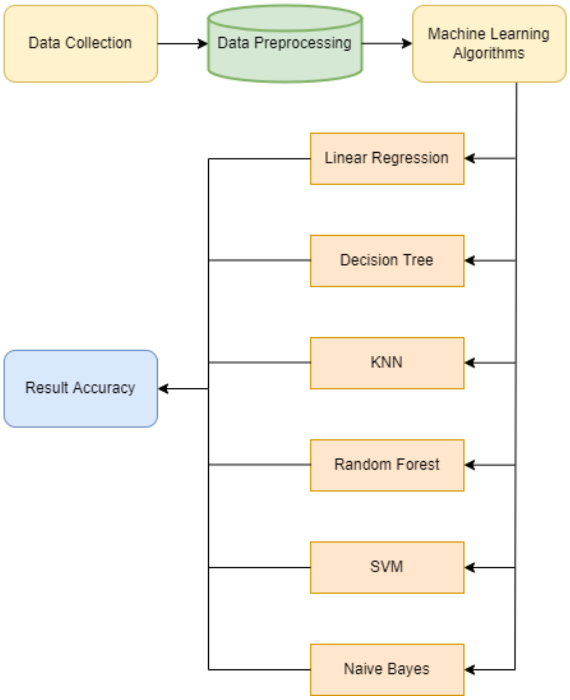
The MSE can be mathematically defined as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is similar to the Mean Standard Error (MSE) in concept, but it has nearly identical characteristics! Take the difference between your model's predictions and the ground truth, multiply it by the absolute value, then average it across the entire dataset to get the MAE.

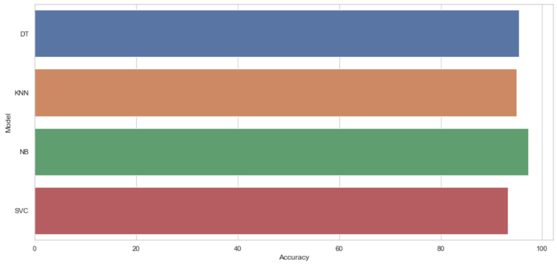
$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$



Result Analysis

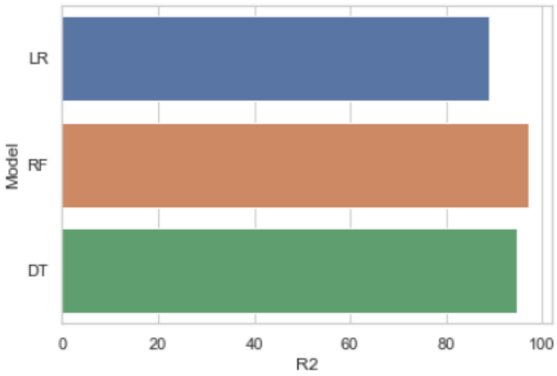
For Crop Recommendation Dataset

Algorithm	Prediction Accuracy
Decision Tree	95.45
KNN	95.00
Naive Bayes	97.28
SVM	93.18



For Crop Yield Prediction Dataset

Algorithm	MSE	MAE	RMSE	R2
Linear Regression	0.107	0.199	0.328	0.89
Random Forest	0.026	0.083	0.163	0.97
Decision Tree	0.0517	0.105	0.227	0.94



Conclusion

From the results obtained from experimenting with different models. Using machine learning techniques like naive Bayes, decision tree and random forest have improved the accuracy of the model drastically when compared to other techniques.

In this project, we've performed machine learning techniques and analysis on two datasets. The first dataset consists of data regarding the classification of the crop. Out of the four machine learning algorithms applied, Naive Bayes had performed better when compared to the others. The second data consists of data regarding yield prediction. Out of the three machine learning algorithms, Random Forest outperformed the other techniques

These algorithms can be used in developing a web application where we could deploy these models for real-life analysis.

References

1. S. Pudumalar et al., "Crop recommendation system for precision agriculture", 2016 Eighth International Conference on Advanced Computing (ICoAC), 2017.

2. A study on various data mining techniques for crop yield prediction", 2017 International Conference on Electrical Electronics Communication Computer and Optimization Techniques (ICEECCOT), 2017.

3. Big data analysis technology application in agricultural intelligence decision system", 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018.

4. Crop recommendation system to maximize crop yield in ramtek region using machine learning", International Journal of

Scientific Research in Science and Technology, vol. 6, no. 1, pp. 485-489, 2019.

5. A Software Model for Precision Agriculture for Small and Marginal Farmers", At the International Centre for Free and Open Source Software (ICFOSS) Trivandrum India, 2013.
6. Comparison of Self Organizing Maps and Sammon's Mapping on agricultural datasets for precision agriculture", International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS), 2015.
7. https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset