CROP PREDICTION USING MACHINE LEARNING-A STEP TOWARDS SELF RELIANCE

—

Smriti Sharma
MTECH. - AI and DS
01902212022

Overview

India is considered as a global agricultural powerhouse by the World. India holds the second largest agriculture land in the world with approximately 179.9 million hectares under cultivation. It is the second largest producer of rice, wheat, cotton, sugarcane, fruits, vegetables and tea.

The Agricultural Sector plays a very important role in the Indian Economy , contributing about 18.8% (FY 2021-22) to the country's GDP . A major strata of India's Population rely on Agriculture for their livelihood. It has been surveyed that approximately 54.6% of India's population is engaged in agriculture.

As the population of India is increasing , so is the demand for food .

Since the Indian farmers are not educated well and afraid of cultivating different crops, crop diversification is the biggest challenge. We still are the largest importer of food products. The farmers have been cultivating the same crops for generations and have a belief in the myth that using more fertilizers and pesticides will yield more production. Instead of soil analysis and cultivating the environmentally suitable diversified crops which will not only help them to earn more but also will meet the diversified food demands of the Indian Population. Many Indian farmers are still using irrationally high quantities of Fertilizers and Chemicals in fields which adversely affects soil, ground water and quality of crops.

Problem Statement

With the increased use of fertilizers, pesticides and other chemicals by farmers for maximizing the yield of crops within a shorter span of time in order to earn more profit results in the production of very high chemically contaminated crops of grains ,fruits and vegetables . These are considered to be one of the key contributors behind various life threatening diseases in human beings and animals. The use of these chemicals/pesticides not only affects humans and animals but also has a significant effect on the environment particularly Groundwater.The use of chemicals/pesticides results in contamination of Groundwater, which will affect the generations to come.

Moreover the high use of fertilizer and pesticides adds to the pocket of farmers.Taking this phase forward and adding to sustainable farming, Our project “Crop Prediction Using Machine Learning- A Step Towards Self Reliance ” aims at :-

1. Limiting the use of fertilizers and pesticides while farming ,
2. Increasing the income of the farmers, by providing them with appropriate crops that are suitable to be grown on their land, depending upon the soil analysis (Crop Prediction) and also providing them about the overall profitability in the entire process (Price Prediction).

This would benefit us in the following manner:-

1. The quality of the crop is maintained automatically , as that crop is grown in the most favorable soil condition based on soil analysis which leads to very limited and restricted use of fertilizers and pesticides.
2. It will yield more profits to the farmer as the cost of buying fertilizers and pesticides is restricted and the labor cost will be reduced significantly.
3. It would also add to the efforts of the government towards the diversification of crops and increase the income of the farmers.
4. Restricted / Limited use of fertilizers and pesticides will not adversely affect the environment, particularly quality of groundwater.
5. Dependability on import will reduce significantly and would help India in achieving the Goal of Self Reliance in the agricultural Sector.
6. Diversified Crop Production, High Yield of crop per acre, Export Quality crops will add up in India’s GDP.

Datasets

1. The dataset used for the research is the Crop Recommendation Dataset which is a class balanced dataset. The dataset consists of different varieties of crops such as wheat ,rice, maize etc. which require different soil and weather conditions for cultivation. The size of the dataset was 2200 rows and 8 columns , the dataset was last updated in 2020 .The dataset consists various features such as (N) ratio of Nitrogen content in the soil, (P) ratio of phosphorus content in the soil, (K) ratio of potassium content in the soil, temperature in degree celsius ,relative humidity in the atmosphere in(percent) , ph value of the soil, amount of rainfall in mm and the target column which is the variety of crop.

   In order to widen up the area of our research we added a few more varieties of crops to our dataset thus increasing the size of the dataset to 3100 rows and 8 columns .The data for (N, P, K , Temperature, Humidity, Ph, Rainfall) for the newly added crops are collected by referring to various Government websites such as Farmer’s Portal.gov.in and Crop Cultivation Guide.

Taking into consideration soil analysis and weather conditions. This dataset helps us to predict the most suitable crop , which is favorable to be cultivated under the given conditions.

LINK- CROP RECOMMENDATION DATASET| KAGGLE

2. The second dataset taken into consideration for the purpose of price prediction is Agri Commodity Min, Max, Modal Price Dataset. The dataset was last updated in 2021. The size of the dataset is (62.31MB). It consists of 348 Commodities prices market wise. This dataset consists of the Name of the Commodity ,State , District, Market, Minimum Price (Quintal => Kg Converted) ,Maximum Price (Quintal => Kg Converted) , Modal Price (Quintal => Kg Converted), Type of Crop,Date on which the data is obtained

LINK-🌾 AGRI COMMODITY MIN , MAX ,MODAL PRICE DATA | KAGGLE

Data Cleaning

Data preparation:

- **In Excel**

  Values of N-P-K of a few crops are added.

- **In Python**

  Data is uploaded via the Pandas read_csv() function and returns a new Dataframe with the data and labels from the file data. csv.

  NULL values are identified for the value column and rows containing NaN in the value columns are dropped.

Exploratory Data Analysis (EDA)

Data analysis utilizing visual methods is called exploratory data analysis (EDA).
With the aid of statistical summaries and graphical representations, it is used to identify trends, patterns, or to verify hypotheses.
Before beginning the modeling work, EDA is used to see what the data can tell us. Finding significant data qualities by looking at a column of numbers or an entire spreadsheet is difficult. Insights can be gained by looking at plain numbers, but doing so can be time-consuming, dull, or overwhelming. Techniques for exploratory data analysis have been developed to help with this. EDA provides the context necessary to create an acceptable model for the problem at hand and to accurately understand its results, making it an essential step to take before delving into machine learning or statistical modeling.

a. Descriptive EDA

- Average Value Of Each Feature in the DataSet is shown.
- Min , Max , Average requirement of N , P , K , Temperature , Humidity , Ph , Rainfall for each crop is shown.
- Listing the average value of a particular feature for all the crops is shown.
- Separating crops on the basis of Average value of the features (N , P , K , Temperature , Humidity , ph , Rainfall) is shown.
- Correlation between the Features is shown using subplots.
- Density plots for all Features is shown.

b. Visualization EDA

1. Bar graph

Information is graphically represented in a bar graph. To represent value, it makes use of bars that stretch to various heights.
Vertical bars, horizontal bars, grouped bars (several bars that compare values in a category), and stacked bars can all be used to construct bar graphs (bars containing multiple types of information).

- Pictorial Representation to illustrate that the dataset is Balanced is shown by bars.

2. Histogram

The underlying frequency distribution (shape) of a set of continuous data can be found and displayed using a histogram, which is a graphic. As a result, the data can be examined for its underlying distribution (such as the normal distribution), outliers, skewness, etc.

- N-P-K values comparison between crops is done.
- Histogram for Temperature-Humidity-Rainfall values comparison between crops is done.

3. Box Plots

A box and whisker plot, often known as a box plot, shows a data set's five-number summary. The minimum, first quartile, median, third quartile, and maximum make up the five-number summary.

- Box plot for all the Features is made separately.

4. Scatter Plot

A scatter plot is a two-dimensional data visualization in which the values of two different variables are represented by dots, with one variable shown along the x-axis and the other along the y-axis.

- PairPlots One vs All is done between all the features.
- Scatter plots for all the Crops to find the underlying trends are done.

Feature Selection and Train-Test-Split

The main objective of our project is to predict the type of crop that is most suitable to be cultivated on the basis of soil analysis and weather conditions.

Hence the feature vector would consist of features such as nitrogen, phosphorus, potassium content in the soil, pH level of the soil and weather conditions such as temperature, relative humidity and the amount of rainfall required by each crop for cultivation.

The classification techniques used in our research consists of two phases: the Training phase and the Testing phase. 70 percent of the data from the dataset is used for training of the model and 30 percent of the data is used for testing purposes.Hence , by using the train_test_split model selection from the sklearn library , we split our dataset into 2 parts, which are represented as X_train, Y_train, X_test, Y_test.

Logistic regression

Logistic regression is a statistical analysis method that predicts a binary outcome, such as yes or no, based on previous observations from a data set.A logistic regression model forecasts a dependent data variable by examining the correlation between one or more already present independent variables.

The technique of logistic regression has grown in significance in the field of machine learning. It enables machine learning algorithms to categorize incoming input based on previous data. The algorithms get more accurate at predicting classes within data sets when new pertinent data is added. A mathematical equation that roughly approximates the relationships between the many variables being modeled is what regression models essentially represent or embody. Machine learning models use input and output data to train on, and then use fresh data to forecast the outcome.

For applying the logistic regression model to our dataset, we first imported the logistic regression model from the sklearn library by using

Import LogisticRegression from sklearn.linear_model

After importing the model, we trained the model using the training dataset. Once the training was completed, the model was allowed to predict the values for the test data, which was denoted by Y_predict.

After receiving the predicted values for the test data , we now wanted to check how accurate our model is. In order to achieve this goal, we made use of evaluation metrics such as accuracy , confusion matrix, and AUC-ROC Curve.

The accuracy of the model was calculated by checking the true values(Y_true) against the predicted values (Y_predict) for the testing dataset. As a result, we achieved a 95% accuracy for logistic regression.

The confusion matrix, which is another evaluation metric that is used to measure the performance of our machine learning model, It consists of four parts

TP (True Positive): The actual class is positive and the predicted class is also positive.

TN (True Negative): The actual class is negative and the predicted class is also negative.

TP (False Positives): The actual class is negative but the predicted class is positive.

TP (False Negative): The actual class is positive but the predicted class is negative.

Precision is defined as how much were correctly classified as positives to all the positives = TP/(TP + FP).

Recall is defined as the ratio of how much were correctly identified as positives to how much were actually positive = TP/(TP +FN).

F1 score is the harmonic mean of Precision and Recall . = 2* Precision* Recall/(Precision + Recall)

Hence, we plotted the confusion matrix for the logistic regression model. The numbers present on the diagonal depict the number of data points that were correctly classified.

The AUC-ROC curve is another performance metric which gives us the summary of our model . It is plotted between TPR (True Positive Rate) and FPR (False Positive Rate) or Precision vs. Recall . ROC stands for Receiver Operator Characteristic, which is used to compare different classifiers. AUC stands for Area Under the Curve . A model with an AUC =1 indicates that the model is perfectly able to distinguish between the positive class and the negative class, and hence the model is good . A model with 0.5< AUC <1 has a high chance of being able to distinguish positive class values from negative class values, hence the classifier is considered good enough . When AUC = 0.5, then the classifier is not able to distinguish between positive and negative class points. Either the classifier is predicting a random class or a constant class for all the data points. This model is not considered good enough. A model with an AUC = 0 is the worst model as it cannot distinguish between the classes and, in fact, it predicts the positive class as negative and the negative class as positive.

On plotting the AUC-ROC curve for the Logistic Regression Model we got to know that each class has an AUC >0.99 . Hence our model is considered really good as it has high probability of correctly predicting the classes

Support Vector Classifier

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method.

For applying the SVC model to our dataset, we first imported the support vector classifier model from the sklearn library by using

 from sklearn.svm import svc

After importing the model, we trained the model using the training dataset. Once the training was completed, the model was allowed to predict the values for the test data, which was denoted by Y_predict.

After receiving the predicted values for the test data , we now wanted to check how accurate our model is. In order to achieve this goal, we made use of evaluation metrics such as accuracy , confusion matrix, and AUC-ROC Curve.

The accuracy of the model was calculated by checking the true values(Y_true) against the predicted values (Y_predict) for the testing dataset. As a result, we achieved a 97% accuracy for svc.

Hence, we plotted the confusion matrix for the support vector classifier model. The numbers present on the diagonal depict the number of data points that were correctly classified.

On plotting the AUC-ROC curve for the SVC Model we got to know that each class has an AUC >0.98 . Hence our model is considered good as it has a high probability of correctly predicting the classes.

Decision Tree Classifier

By using a greedy search to find the ideal split points inside a tree, decision tree learning uses a divide and conquer technique. When most or all of the records have been classified under distinct class labels, this splitting procedure is then repeated in a top-down, recursive fashion. The intricacy of the decision tree plays a significant role in determining whether or not all data points are categorized as homogeneous sets. Pure leaf nodes, or data points belonging to a single class, are easier to obtain in smaller trees. It gets harder to preserve this purity as a tree gets bigger, which typically leads to too little data falling under a particular subtree. Data fragmentation is the term used to describe this situation, and it frequently results in overfitting.

For applying the Decision Tree model to our dataset, we first imported the Decision Tree classifier model from the sklearn library by using

```
 from sklearn.tree import DecisionTreeClassifier
```

After importing the model, we trained the model using the training dataset. Once the training was completed, the model was allowed to predict the values for the test data, which was denoted by Y_predict.

After receiving the predicted values for the test data , we now wanted to check how accurate our model is. In order to achieve this goal, we made use of evaluation metrics such as accuracy , confusion matrix, and AUC-ROC Curve.

The accuracy of the model was calculated by checking the true values(Y_true) against the predicted values (Y_predict) for the testing dataset. As a result, we achieved a 98% accuracy for Decision Tree Classifier..

Hence, we plotted the confusion matrix for the Decision Tree classifier model. The numbers present on the diagonal depict the number of data points that were correctly classified.

On plotting the AUC-ROC curve for the Decision Tree Classifier Model we got to know that each class has an AUC >0.90 . Hence our model is considered good as it has a high probability of correctly predicting the classes.

Random Forest Classifier

A very popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance.

Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead, then depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. Higher accuracy and overfitting are prevented by the larger number of trees in the forest.

Reasons to use Random Forest:

1.In comparison to other algorithms, it requires less training time.

2.Even with the enormous dataset, it operates effectively and predicts the outcome with a high degree of accuracy.

3.When a significant amount of the data is absent, accuracy can still be maintained.

Method to Construct Random Forest:

First, N decision trees are combined to generate the random forest, and then predictions are made for each tree that was produced in the first phase.

Steps for decision tree:

Step 1: Pick K data points at random from the training set.

Step 2: Construct the decision trees linked to the chosen data points (Subsets).

Step 3: Select N for the size of the decision trees you wish to construct.

Step 4:Repeat steps 1 and 2 in step 4.

Step 5: Assign new data points to the category that receives the majority of votes by looking up each decision tree's predictions for the new data points.

Implementation of the Random Forest algorithm in Python

The following are the steps for implementation:

1. The pre-processing of data.

2.Aligning the Training set with the Random Forest algorithm.

3. Foreseeing the test outcome.

4. Evaluate the result's correctness (Creation of Confusion matrix).

The accuracy of the model was calculated by checking the true values(Y_true) against the predicted values (Y_predict) for the testing dataset. As a result, we achieved a 99% accuracy for Decision Tree Classifier..

Random Forest's advantages.

Both classification and regression tasks can be handled by Random Forest.

It is able to handle big datasets with lots of dimensions.

It improves the model's accuracy and avoids the overfitting problem.

Related Research Papers

- **Analysis of agricultural crop yield prediction using statistical techniques of machine learning**

  Authors - Janmejay Pant , R.P.Pant, Manoj Kumar Singh , Devesh Pratap Singh HimanshuPant

  LINK- ANALYSIS OF AGRICULTURAL CROP YIELD PREDICTION USING STATISTICAL TECHNIQUES OF MACHINE LEARNING

- **Crop Price Prediction Using Random Forest and Decision Tree Regression:-A Review**

  AUTHORS- Manik Rakhra , Priyansh Soniya, Dishant Tanwar, Piyush Singh, Dorothy Bordoloi, Prerit Agarwal, Sakshi Thakkar, Kapil Jarath, Neha Verma

  LINK- CROP PRICE PREDICTION USING RANDOM FOREST AND DECISION TREE REGRESSION: -A REVIEW

- **Crop yield prediction using machine learning: A systematic literature review**

  AUTHORS- Thomas van Klompenburg Ayalew, Kassahun Cagatay, Catal Ayodele

  LINK - CROP YIELD PREDICTION USING MACHINE LEARNING: A SYSTEMATIC LITERATURE REVIEW

- **Classification of crop based on macronutrients and weather data using machine learning techniques**

  AUTHORS- Ritesh Dash, Dillip KuDash , G.C.Biswal

  LINK - CLASSIFICATION OF CROP BASED MACRONUTRIENTS AND WEATHER DATA USING MACHINE LEARNING TECHNIQUES

LINK FOR PRESENTATION

CROP PREDICTION USING MACHINE LEARNING- A STEP TOWARDS SELF RELIANCE