ML Project Interim Report

Crop Recommendation using ML

1. Abstract

India's economy is majorly dependent on agriculture. Due to the increasing human population, it is more prominent to utilize agricultural lands as efficiently as possible so as to get maximum efficiency. Choosing right crop for an agricultural land is very crucial. Due to the limited knowledge, it is sometimes difficult for the farmers to choose right crop for their field.
Due to the wrong choice of crop, farmers suffer a lot. It becomes difficult for them to get any profit.
If the profit is less, it will be impossible for farmers to expand their agriculture.
This machine learning project will helps the farmers choose the crop based upon climatic and soil conditions which can be produced in their lands more efficiently.

2. Introduction

Precision agriculture is in trend nowadays. Precision agriculture is a modern farming technique that uses the data of soil characteristics, soil types, crop yield data, and weather conditions and suggests the farmers with the most optimal crop grow in their farms for maximum yield and profit. This technique can reduce crop failures and help farmers make informed decisions about their farming strategy.

This machine learning project's objective is to find the best suitable crop for agricultural land by learning from the past yielded crops. Various factors are considered for determining the best yielding crop, such as the climate, which includes rainfall, temperature, and the soil contents such as the pH level, nutrient content such as N, P, K of the soil, and more. During this project, several machine learning algorithms are applied, and performance comparison is made between them. Also if time permits using the above-mentioned model, we will implement another model that will predict the yield of the previously predicted crop.

3. Literature Review

3.1 Prediction of Crop Yield Using Regression Techniques

Aditya Shastry, H.A. Sanjay and E. Bhanusree
Nitte Meenakshi Institute of Technology, Bangalore, India
This study explains the experiments carried out on wheat, cotton and maize data sets using quadratic, pure quadratic, linear, polynomial, generalized linear regression and stepwise linear regression models. It also compares the results obtained from them. Accuracy of these prediction models are measured using (R°), Root Mean Square Error (RMSE) and Mean Percentage Prediction Error (MPPE). The model which gives lower Root Mean Squared Error, Prediction Error and higher R° statistics is considered to be the best model for the crop yield prediction.

3.2 PREDICTING YIELD OF THE CROP USING MACHINE LEARNING ALGORITHM

P. Priya, U.Muthaiah, M.Balamurugan

In this study, the data sets considered are rainfall, perception, production, temperature to construct random forest, a collection of decision trees by considering two-third of the records in the datasets. These decision trees are applied on the remaining records for accurate classification. The resultant training sets can be applied on the test data for correct prediction of crop yield based on the input attributes. Random Forest algorithm was used to study the performance of this approach on the dataset. The advantage of random forest algorithm is that overfitting is less of an issue, unlike decision tree machine learning algorithms. There is no need of pruning the random forest. Random Forest machine learning algorithms can be grown in parallel. The Results show that we can attain an accurate crop yield prediction using the Random Forest algorithm. Random Forest algorithm achieves the largest number of

crop yield models with the lowest models. It is suitable for massive crop yield prediction in agricultural planning. This makes the farmers take the right decision for the right crop such that the agricultural sector will be developed by innovative ideas.

4. Datasets

The data used in this project is made by augmenting and combining various publicly available datasets of India like weather, soil and others. This data is relatively simple with very few but valuable features, unlike the complicated features affecting the crop's yield.
The data have Nitrogen, Phosphorous, Potassium and pH values of the soil. It also contains the humidity, temperature and rainfall required for a particular crop.

4.1 Features

The features for our model are:

[1] Rainfall in mm
[2] Temperature in Celsius
[3] Humidity
[4] pH level of soil
[5] Nitrogen (N) level of soil.
[6] Phosphorous (P) level of soil.
[7] Potassium (K) level of soil

4.2 Dataset Description

Our output label is multiclass classification.

The dataset consists of 22 different class labels.

Our dataset consists of 69,718 data samples

	N	P	K	temperature	humidity	ph	rainfall
count	69718.000000	69718.000000	69718.000000	69718.000000	69718.000000	69718.000000	69718.000000
mean	0.298951	0.318797	0.256143	0.163011	0.435709	0.041202	0.591565
std	0.205225	0.170283	0.183024	0.058039	0.155403	0.013535	0.207400
min	0.000000	0.025451	0.032393	0.032439	0.093348	0.016189	0.127630
25%	0.126465	0.160105	0.141933	0.125016	0.320216	0.030584	0.421744
50%	0.248396	0.316834	0.185572	0.148612	0.437924	0.040944	0.582216
75%	0.466877	0.471069	0.286924	0.199804	0.567565	0.049222	0.782401
max	0.776165	0.714726	0.783144	0.359141	0.800347	0.104800	0.926167

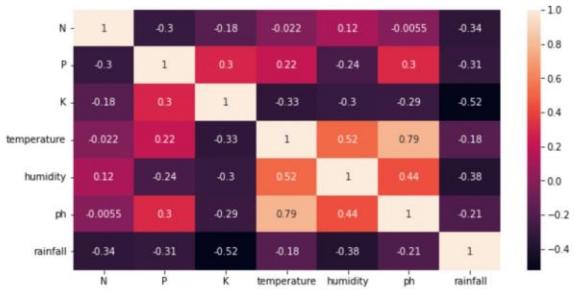
.

```
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   N            69718 non-null  float64
 1   P            69718 non-null  float64
 2   K            69718 non-null  float64
 3   temperature  69718 non-null  float64
 4   humidity     69718 non-null  float64
 5   ph           69718 non-null  float64
 6   rainfall     69718 non-null  float64
 7   label        69718 non-null  object
dtypes: float64(7), object(1)
```

4.3 Preprocessing

4.3.1 Feature Selection and filtering / removed out the unnecessary from the dataset.
4.3.2 Checked for null values and Nan values and transformed data for that values
4.3.3 Checked for duplicate samples in the dataset
4.3.4 Checked for correlation among the features in the dataset using heat maps and pairwise plots
4.3.5 Normalization of the features to change the values of numeric columns in the dataset to a common scale

```
array(['rice', 'maize', 'chickpea', 'kidneybeans', 'pigeonpeas',
       'mothbeans', 'mungbean', 'blackgram', 'lentil', 'pomegranate',
       'banana', 'mango', 'grapes', 'watermelon', 'muskmelon', 'apple',
       'orange', 'papaya', 'coconut', 'cotton', 'jute', 'coffee'],
      dtype=object)
```



5. Methodology

5.1 Libraries Used

5.1.1 Pandas
5.1.2 Numpy
5.1.3 Seaborn
5.1.4 Matplotlib
5.1.5 Sklearn

5.2 Steps followed

5.2.1 Data Preprocessing
5.2.2 Displaying data information and description
5.2.3 Feature selection by removing unnecessary columns
5.2.4 The dataset is supervised learning.

5.3 Relational Analysis

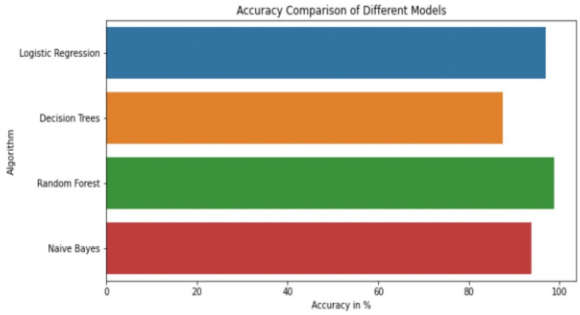
5.3.1 Correlation
5.3.2 Plotting Heat Map
5.3.3 Plotting Pairwise Plot grid
5.3.4 Plotting Relational Plot
5.3.5 Data Distribution Plot
5.3.6 Plotting mean value for each crop type

5.4 Model Training

5.4.1 Logistic Regression
5.4.2 Naïve Bayes Classifier with Cross Validation
5.4.3 Decision Trees
5.4.4 Random Forests

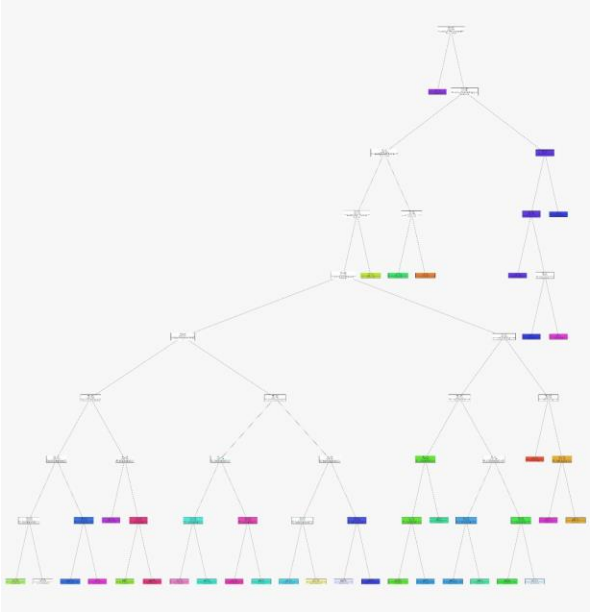
6. Results and Analysis

6.1 In the dataset we don't have any null or Nan values as reported data exploration part.
6.2 In the dataset no duplicate entries/repeated data were present
6.3 On calculating Correlational Heat Map we can observe that the features are linearly dependent and are not related to each other
6.4 Listed different classes for our classification task
6.5 Done Relational Analysis
6.6 Analyzed the learning involved in our classification task as Supervised Learning
6.7 Analyzed accuracies in case of different models such as Naïve Bayes, Logistic Regression, Decision Trees



['Logistic Regression', 'Decision Trees', 'Random Forest', 'Naive Bayes']
[97.04054312488047, 87.54541977433544, 98.71390323197552, 93.70816599732262]

Decision Tree



7. Conclusion

7.1. Learning from Project

7.1.1 Agriculture is the backbone of Indian Economy. ML models can be utilized in farming sector to improve agricultural yield.
7.1.2 Finding a good dataset.
7.1.3. Importance of correct dataset. Machine learning model is as good as our chosen dataset.
7.1.4. Feature selection. Which features are needed and which are not in our ML Model.
7.1.5 Displaying the information about our data in code (Number of classes, type of each feature, etc.).
7.1.6 Data preprocessing like checking null values, checking for duplicate samples and normalization of data .

7.1.7 Relationship analysis like finding correlation between different features.
7.1.8 Finding mean value of each feature for each label in our code.
7.1.9 Get to know how different model are performing on the same dataset

7.2 Work Left

7.2.1. Analysis based upon previously modeled data
7.2.2. Classification and Data Training on some other models
7.2.3. Applying some more different models.
7.2.4 If time permits going to implement another model involving results of this model for training of other model

7.3 Contribution

Harsh:
Finding dataset, data extraction, dataset cleaning, making presentation, Logistic Regression, Random Forest

Nitin:
Project Idea, EDA, data preprocessing, plotting graph, making presentation, Normalization of data, Analysis of Random Forest

Dhananjay:
Project Idea, data extraction, dataset cleaning, report writing, Naïve Bayes, Decision Trees

Ishaan:
Finding dataset, EDA, plotting graphs, report writing, relationship analysis.

References

http://www.ijstr.org/final-print/jan2020/Design-And-Implementation-Of-Crop-Yield-Prediction-Model-In-Agriculture.pdf

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9214190