Machine Learning Assignment

Crop Recommendation Project

The Capstone Project

Saumya Gunawardana
#196

Introduction

When it comes to farming, the best suitable crop for different climate conditions vary according to the parameters/indicators related to the climate. Over a period, agriculture specialists have done research on factors that affects the decision of selecting the proper crop. Soil characteristics, soil types, crop yield data, weather condition etc. are the factors that a farmer must consider prior to selecting the crop to have a maximum yield.

A set of collected data can be used to support in making this decision. Therefore, it is possible to use machine learning to reduce the crop failures and to help the farmers to take informed decisions about their farming strategy.

Data

The dataset that was selected for the project contains data gathered on the best yielding crops for a certain climate condition. The data set consists of 2200 records that contain below attributes.

N : Nitrogen level of soil

P : Phosphorus level of soil

K : Potassium level of soil

Temperature : temperature of the environment

Humidity : humidity of the environment

Ph : pH level of soil

Rainfall : rainfall to the region

Label : recommended crop based on the conditions

Each of these attributes are represented in different variable types as listed below.

Column #	Column Header	Data type
0	N	int64
1	P	int64
2	K	int64
3	temperature	float64
4	humidity	float64
5	ph	float64
6	rainfall	float64
7	label	object

The data set recommends a set of crop types namely;

- rice
- maize
- chickpea
- kidneybeans
- pigeonpeas
- mothbeans
- mungbean
- blackgram
- lentil
- pomegranate
- banana
- mango
- grapes
- watermelon
- muskmelon
- apple
- orange
- papaya
- coconut
- cotton
- jute
- coffee

Methodology

After analyzing data, it was identified that the data set can be used to develop a machine learning model to predict the best yielding crop based on the given parameters.

Data Preprocessing

Upon loading the data set below listed preprocessing activities were done prior to using data to the machine learning model.

- Explored the data set to check how many records are there for each crop type
- All the spaces were removed from data labels to avoid any confusions due to data inconsistency
- Since there were no empty/blank entries, it was not required to remove blank entries

Since the selected data set was a very much cleansed one, not much of effort was taken in order to prepare and rearrange the data.

Machine Learning Models

Since the decision making is to be trained using an already available dataset, the machine learning approach to be used is **Supervised Learning**.

The total dataset was split into two sets simply to train the model and to validate the model. Upon segregating the data was fed to below machine learning models.

- Logistic Regression Model
- Decision Tree Classifier
- Random Forest Classifier

Comparing Models

Upon applying the models, their performance was tested based on several parameters, namely;

- Accuracy
- Precision
- Recall
- F1 Score

Results

Achieved scores are depicted in the below table.

	Model	accuracy	precision	recall	f1_score
0	Logistic Regression	0.957576	0.958339	0.957576	0.957889
1	Decision Tree	0.869697	0.828091	0.884874	0.831863
2	Random Forest	0.992424	0.994172	0.991883	0.992312

Conclusion

By referring to the achieved scores, it was decided that the **Random Forest Classifier** is the best suitable model to be used on the selected data set. Therefore, it was decided to use Random Classifier Model on the data set in predicting the best yielding crop based on the given parameters.

The developed model was applied by using a set of random input parameters and the model predicted the best yielding crop for each set of input parameters by applying the trained model.

Discussion

The selected data set of the project is comparatively a clean data set where I did not require to apply much effort on data cleansing. In addition, the given data set had only the essential required data

fields in decision making. Therefore, feature engineering effort and also the data restricting efforts were also very minimal. But when it comes to real world scenarios it was understood that it is required to apply a considerably huge effort in cleansing data prior to be applied in a model. Further, feature engineering does a critical role in shaping data prior to be used to train a model.

But learning gained by carrying out the project enhanced my practical knowledge of machine learning so that I can improve on other aspects when it comes to practical application of my learnings into organizational environment.

Further the given crop database had a proper data set which provides a higher accuracy levels. Therefore machine learning model generated accurate predictions. But when it comes to real world data, it is challenging to compare models and to decide on a model if the selected data set has inaccurate information which diverts the model towards wrong predictions.

Further, the model could be developed with an interface to provide predictions upon inserting the required parameters, so that it can be shared as a practical model to be used in farming. But I have not gone to that extent due to the time limitation and would like to explore these aspects in my future projects.