Crop Prediction Model using Random Forest Classifier

By Aditya Rawat

Abstract

This research introduces you to a Data Science System which is designed to assist farmers in selecting the most suitable crop for cultivating based on the details of the soil. Example – pH level, Nitrogen level, Phosphorus level, and other important metrics. We used Random Forest Classifier Algorithm to provide tailored crop recommendation based on soil analysis.

A comprehensive dataset of soil detail and corresponding successful crop yields was taken from the kaggle – Crop Recommendation Dataset. The dataset was used to train the various machine learning models, the best performing model that was Random Forest Classifier was selected based on its accuracy and reliability in predicting optimal crop.

The implementation of this AI/DS model represents a significant advancement in precision agriculture practices, enabling farmers to make informed, data-driven decisions that enhance crop yield and sustainability.

Soil degradation is the decline in soil characteristic originated by its inappropriate use, typically for agricultural, pastural, industrial or urban causes (Johnson and Lewis, 1995). It is a severe universal ecological crisis and may be aggravated by weather change. Soil degradation is one of the most important threats facing mankind which not only weakens the productive capability of an ecosystem but also affects overall climate. [1]

Keywords: Artificial Intelligence, Data Science, Random Forest Classifier

Introduction

Agriculture is the primary source of trade which forms the backbone in many different countries. For the improvement of production of crop, soil moisture data are used. Soil moisture nutrients are essential for better enhancement of crop growth and plant tissues. Crop prediction depends on important parameters of a region like – many harvesting techniques, weather conditions (such as cloud, rainfall, temperature, humidity etc.), soil moisture type (such as clay, sandy, saline soil, etc.), soil composition (such as pH value, potassium, nitrogen, phosphorous, etc.) and many more. The soil composition data can be analysed by using various machine learning algorithms. The system is created using machine learning algorithm and it predicts appropriate crops to the farmers depending on the quality of the soil in the region. [2]

In this research paper, we used Random Forest Classifier machine learning algorithm which helps to predict crop yield in agriculture that can help framers. The model predicts and suggests most accurate crop based on the soil attributes. This prediction will help farmer to predict the yield of the crop before cultivating and thus help them to make appropriate decisions. We used the following soil details to predict the crop – which needed to grow.

Sno.	N	P	K	temperature	humidity	ph	rainfall	label
1	90	42	43	20.87974371	82.00274423	6.502985292	202.9355362	rice
2	71	54	16	22.61359953	63.69070564	5.749914421	87.75953857	maize
3	40	72	77	17.02498456	16.98861173	7.485996067	88.55123143	chickpea
4	23	72	84	19.02061277	17.13159126	6.920251378	79.92698081	chickpea
5	13	60	25	17.13692774	20.59541693	5.68597166	128.256862	kidneybeans
6	9	66	21	30.11812084	34.13307843	5.719889876	157.0858232	pigeonpeas

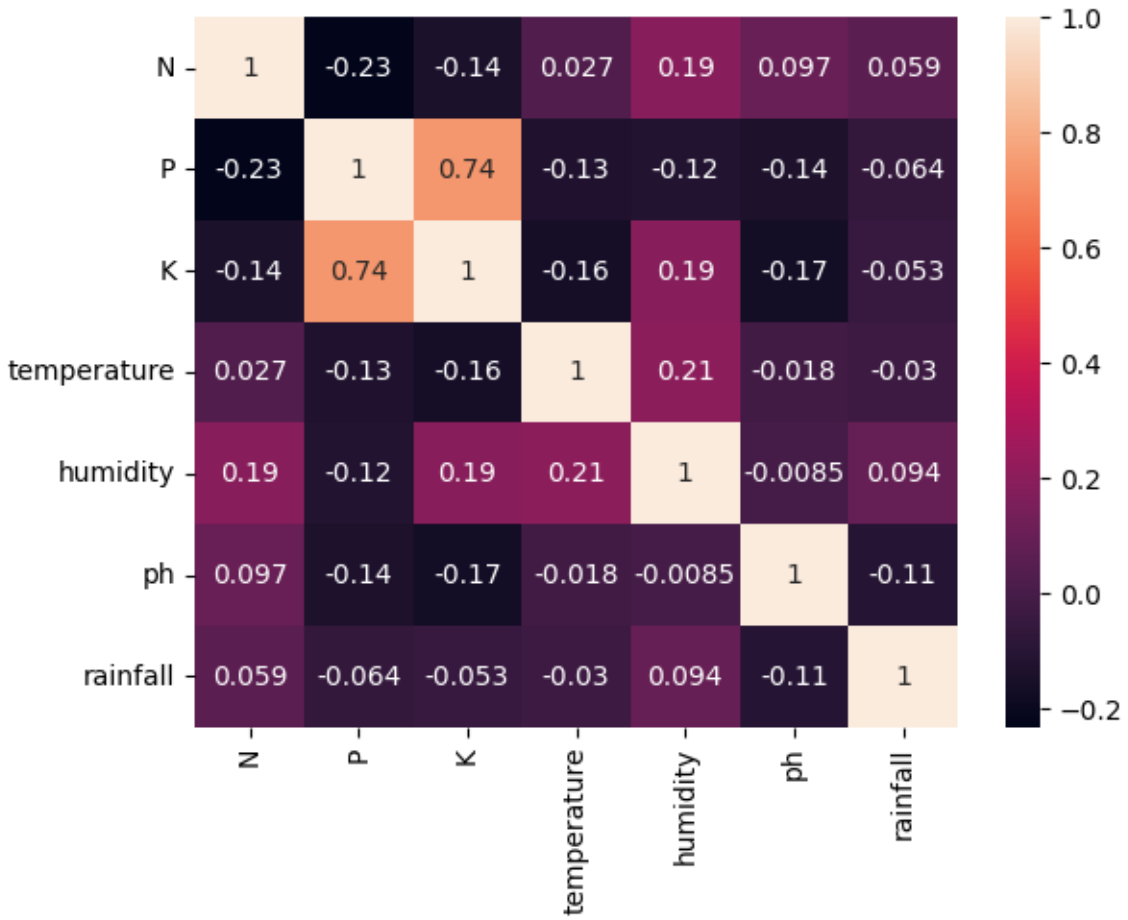


Fig 1. Correlation Heat map of the data

Project Vision

The vision of the project was to revolutionize crop selection process in agriculture by making them more precise, efficient and sustainable. By using the power of artificial intelligence and data science, we aim to empower farmers with a tool, which will help them by providing accurate and personalize crop recommendation based on the soil analysis.

a. Enhance farmers' ability to make informed decision that increase productivity and sustainability.

b. Optimizing resource use, such as water, fertilizers and soil to reduce waste and environmental impact.

c. Increase crop yield by matching crops to the most suitable soil conditions.

Underlying Technologies

1. Python: we used python as the major programing language. It is used to train the Machine Learning Algorithm i.e. Random Forest Classifier and other pre-processing methods.

2. Scikit-learn: It is used to access the pre-processing features such as MinMaxScaler (for normalizing the data) and StandardScaler (for standardization the data). It was also used to train the Random Forest Classifier Model using train_test_split.

3. Flask: we used Flask web frame work to connect the Crop Prediction Model with the HTML website.

Code Design

1. Splitting data into Train and Test:

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)

Here 'train_test_split' function splits the dataset into 80% training and 20% testing set, because the 'test_size' is set to 0.2 Training set of the data is used to train the machine learning model. The model learns the pattern and relationship in data

2. Normalization:

mx = MinMaxScaler()

mx.fit(X_train)

X_train = mx.transform(X_train)

X_test = mx.transform(X_test)

Normalization is used in data pre-processing to adjust the values of numeric features in a dataset to a common scale, without distorting differences in the ranges of values. [3]

Here, we used 'Min-Max Normalization' technique, it is used to rescales the data to a fixed range, usually of range[0,1].

3. Standardization:

sc = StandardScaler()

sc.fit(X_train)

X_train = sc.transform(X_train)

X_test = sc.transform(X_test)

Standardization is a data preprocessing technique that transforms data tohavea mean of 0 and a standard deviation of 1. The aim is to achieve a consistent scale across features, making them comparable and improving the performance of certain algorithms. [4]

4. Training the Machine Learning Model:

randclf = RandomForestClassifier()

randclf.fit(X_train, y_train)

y_pred = randclf.predict(X_test)

Random Forest Classifier is a popular and powerful ensemble supervised classification method. Due to its superior accuracy and robustness, and some ability to offer insights by ranking of its features, Random Forest Classifier has effectively been applied to various machine learning applications, including this project. [5]

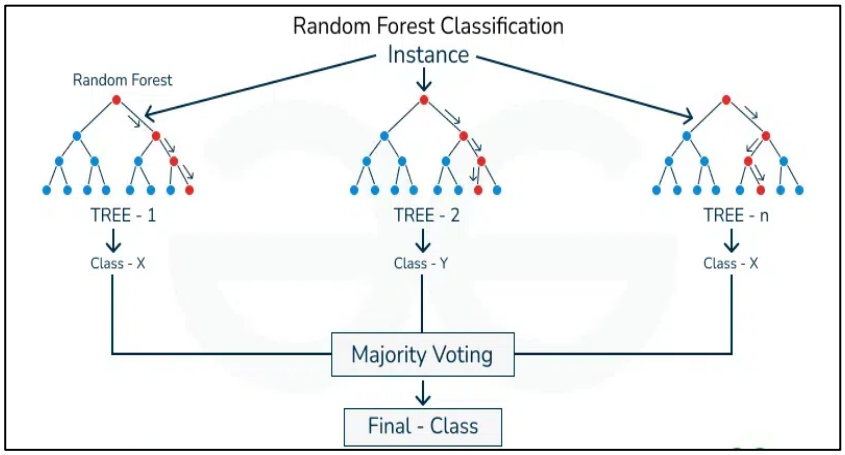


Fig 2. Process of Random Forest Classification

Accuracy

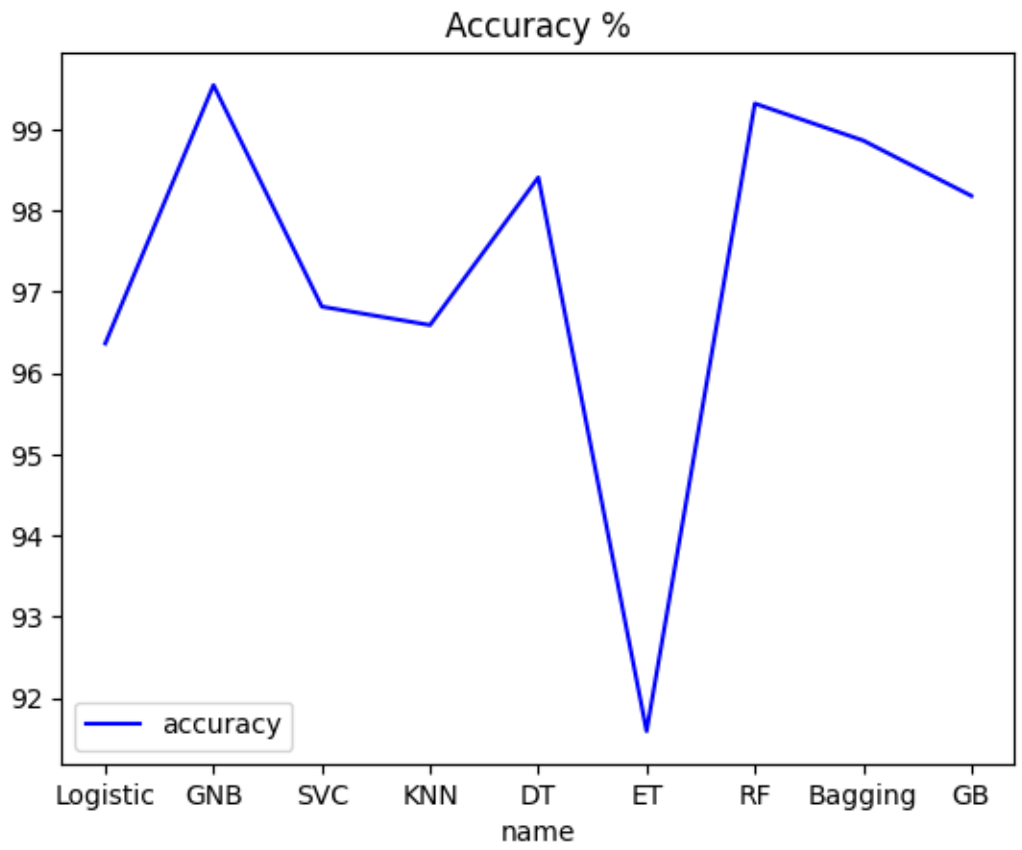


Fig 3. Contrasting Analysis of Machine Learning Techniques

Different types of Machine Learning Models were used to predict the crop based on the soil composition. We used Logistic Regression, Gaussian Naïve Bayes, SVM, KNN, Decision Tree, Random Forest Classifier etc.

Based on the Graph Plotting we can see that Gaussian Naïve Bayes and Random Forest Classifier have the highest peak in the graph. We used Random Forest Classifier in predicting the crop as it is very simple to understand and use.

Process Workflow

Crop Prediction using the following steps:

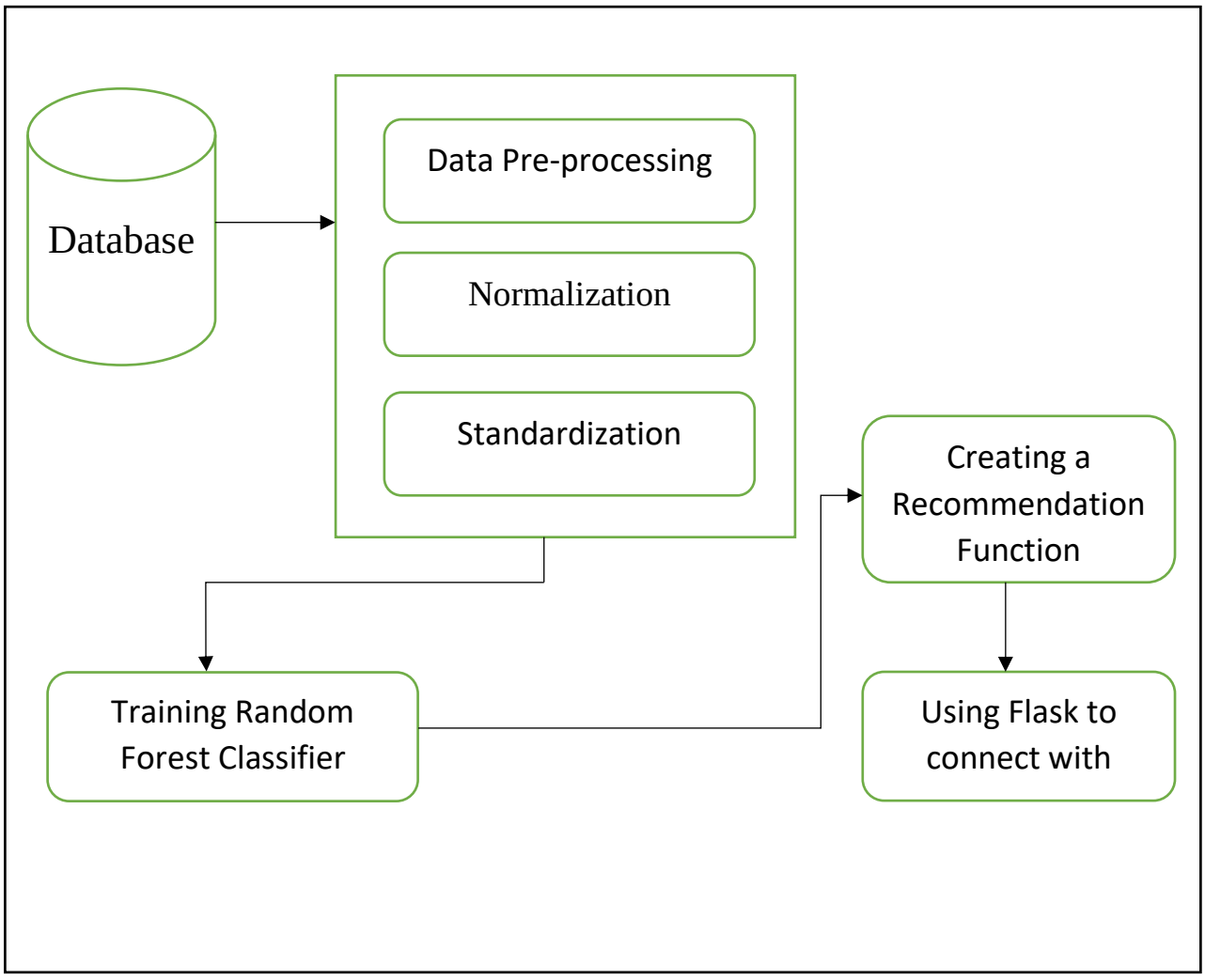
I. Obtaining the data from the database and performing all the pre-processing steps like:
   a. Importing essential libraries
   b. Checking for missing value and outliers
   c. Splitting dataset into training and test sets

II. Normalizing the data, using MinMaxScaler to Normalize the data between [0,1]

III. Standardization: converting the mean value of data to 0 and standard deviation to 1.

IV. Training the Random Forest Classifier Machine Learning Model, to predict the Crop.

V. Connecting the website with the program using Flask web framework.



Conclusion

Machine learning technologies are frequently used for Crop Prediction. In this study, a variety of machine learning techniques are compared for the crop prediction based on soil composition, including Logistic Regression, GNB, SVM, KNN, Decision Tree, Random Forest, etc. on the soil - crop Dataset. The Random Forest Classifier, which obtained an accuracy of 99.31 percent, is the best method. This suggests that the Random Forest Classifier is the best method for detecting Crop Prediction using Soil Composition.

References

[1] Alam, Afroz. (2014). Soil Degradation: A Challenge to Sustainable Agriculture. International Journal of Scientific Research in Agricultural Sciences. 1. 50-55. 10.12983/ijsras-2014-p0050-0055.

[2] T R, Mahesh & G, Sindhu. (2023). Prediction of Crop Yield Based-on Soil Moisture using Machine Learning Algorithms. International Journal of Information Technology, Research and Applications. 2. 33-41. 10.59461/ijitra.v2i1.30.

[3] Shantal, Mohammed & Othman, Zalinda & Bakar, Azuraliza. (2023). A Novel Approach for Data Feature Weighting Using Correlation Coefficients and Min–Max Normalization. Symmetry. 15. 2185. 10.3390/sym15122185

[4] Blessing, Elisha & Klaus, Hubert. (2023). Normalization and Standardization: Methods to preprocess data to have consistent scales and distributions. 2237. 10.

[5] Dragutin Petkovic (), Russ Altman (), Mike Wong (), and Arthur Vigil () Biocomputing 2018. January 2018, 204-215. Improving the explain ability of Random Forest classifier – user centred approach