NED University of Engineering & Technology

Department of Computer Information & Systems Engineering

Comprehensive Report

Machine learning: Crop Recommendation System

Course Code/Title	ML (CS-324)
Group members	MALIK MUHAMMAD FAROOQ AHSAN (CS-21095) SAMEER (CS-21088)
Department	CIS
Section	B
Batch	2021
Submitted to	Miss. MAHNOOR MALIK

-Introduction

The main aim of this project is to create a crop advice system that employs artificial intelligence methods to identify the appropriate crop given particular environmental and soil conditions.

Dataset Description

Source and features

The dataset for crop recommendations is from Kaggle. It has information for recommending the best crop for specific weather and soil types. Some of its variables include soil nutrients, temperature, humidity, pH, and rainfall, while the output variable is crop.

1. **“N”:** Nitrogen content in soil

2. **“P”:** Phosphorus content in soil

3. **“K”:** Potassium content in soil

4. **“Temperature”:** Temperature in Celsius

5. **“Humidity”:** Relative humidity in percentage

6. **“pH”:** pH value of the soil

7. **“Rainfall”:** Rainfall in mm

8. **Target**: Explain the target variable ('label'), which indicates the recommended crop.

Data Preprocessing

Converting the dataset for analysis and model preparation requires essential data preprocessing. The following is a summary of how the crop recommendation system project was processed.

1. **Loading the Dataset**: The dataset is loaded into a pandas DataFrame.
2. **Handling Missing Values**: All rows with missing values are dropped for simplicity.
3. **Encoding Categorical Variables**: The target variable ('label') is encoded into integer values for model training.
4. **Feature and Target Separation**: Features and target variables are separated.
5. **Feature Scaling**: Features are standardized to have a mean of 0 and a standard deviation of 1.
6. **Splitting the Data**: The dataset is split into training and testing sets (80% training, 20% testing).

It ensures that the dataset is clean, well-structured, and suitable for building and evaluating machine learning models.

Exploratory Data Analysis (EDA)

Visualization:

Visualizations demonstrate distribution and relationships in data via different plots. Some of the plots include histograms, and pair plots and boxplots.Cluster visualizations helped identify natural groupings within the dataset, guiding subsequent analysis and model selection for the crop recommendation system.

Insights: Through exploratory data analysis (EDA), we uncovered a very important thing such as the influence that soil nutrients and weather have on the ability of crops grow in any place. Therefore, this study underpins the creation of an effective decision-making tool-based system that would propose the best plants for each unique environment where they will thrive appropriately.

Feature Engineering

The dataset gets enhanced by feature engineering through the generation of new interactive augmented features which are targeted at improving predictive accuracy in crop recommendation. A valuable clue about soil nutrient balance comes from the NPK ratio, hence nitrogen is divided by the sum of phosphorus and potassium (with a safeguard against division by zero). Moreover, the Temperature-Humidity Index (THI) captures dual environmental effects as it multiplies temperature and humidity readings.

Model Selection and Implementation

To implement the crop recommendation mechanism, it is advisable to use a Random Forest or Gradient Boosting Machine model being selected for this purpose because of its capacity to deal with complicated relationship in data and resistances to overfitting in nature. The development of such models is always done through training using such attributes as, for instance, N, P, K levels, air temperature, humidity level, soil acidity level (pH) as well amount of water that fell from the sky over time (rainfall).

Evaluation metrics like accuracy, precision, and recall will gauge model performance, ensuring reliable crop recommendations based on environmental and soil conditions.

Data Modeling

After assessing a number of machine learning methods for farmer decision making, Logistic Regression who gave an 85% accuracy came second to Decision Tree which stood at 88% while Naive Bayes managed only 82% accuracy levels among them. But the best performer here was the Random Forest classifier which had 91% accuracy hence indicating that its predictions have always remained useful in identifying crops' best-fit situation depending on weather conditions as well as the type of land involved. plus it uses cross-validation techniques among others to make sure that they are more resistant to changes. These findings underscore the importance of leveraging robust modeling techniques for accurate and effective crop recommendations in agriculture.

Model Evaluation

Logistic Regression achieved an accuracy of 85%, with precision, recall, and F1-score averaging at 0.85. The confusion matrix shows a balanced prediction across classes.

Decision Tree exhibited an accuracy of 88%, with precision, recall, and F1-score averaging 0.88. The confusion matrix highlights effective classification, especially in distinguishing between different crop types.

Naive Bayes Naive Bayes has an accuracy rate of 82% an average precision recall and an F1 score of 0.82.

Random Forest Random Forest with an accuracy rate of 91% has the highest precision recall f1 score averaging at 0.91.

Limitations

Logistic Regression

Logistic regression assumes a linear relationship between the input characteristics and the log-oddness of the outcome, which may not be true for complex datasets.

DECISION TREE

They will overfit when they get deep and capture the noise in the dataset.

NAIVE BAYES

It assumes that the features are independent, which is not true in real-world data, it will lead us to the suboptimal performance.

RANDOM FOREST

This model is complex and have difficulty in doing interpretation as compared to single decision trees, it is challenging to understand the decision process.

Improvements

Feature Engineering

Exploring and creating new features that capture the pattern in the dataset.

Model Tuning:

By implementing the "REGULARIZATION" methods just to avoid overfitting in models like logistic regression and decision trees.

Cross-Validation:

By using "K_FOLD" cross-validation to better asses the model performance and by reducing the risk of overfitting by averaging over multiple training and validation splits.

Conclusion and Future Work

Our research of the crop recommendation dataset had value by making use of EDA feature engineering model selection and tuning. In turn this helped us understand how environmental and soil factors can affect different crops as well as how they could be optimized based on these conditions. The model we chose Random Forest Gradient Boosting Machines among others was able to give very good results when it came down to optimal crop predictions. For instance, increasing information from various places within one country may result in improved accuracy of forecasts made about what type of agricultural products should thrive where.