Dupli Checker

Date: 12-01-2025

Plagiarism Scan Report

4%
Plagiarism

2%
Exact Match

2%
Partial Match

96%
Unique

Words	1000
Characters	6882
Sentences	86
Paragraphs	82
Read Time	5 minute(s)
Speak Time	7 minute(s)

Content Checked For Plagiarism

AIE425 Intelligent Recommender Systems, Fall Semester24 /25
Productivity- and Season-based Agricultural Crop Recommendation Engine

Student ID, Full Name: 1- Osama Gamal Hamed Ebraheem , B20000007
2-Sohila Mohamed Ali , A20001134

10.1. Explanation of the Process of Data Collection and Data Preprocessing

Data Collection.

The dataset we are using for the Productivity- and Season-Based Agricultural Crop Recommendation Engine project was taken from two sites.

Crop Recommendation Dataset.

The dataset contains N, P, K, temperature, humidity, pH, rainfall, and crop tag. The crop tag defines the suitability of various crops under given conditions.

Machine learning algorithms for crop recommendation systems were sourced from the publicly available repository.

Crop Production Dataset.

This dataset has previous years' records of agricultural productivity such as the quantity of produce, area and season of crop.

I pulled out information from government records and agri-statistics websites to create a model that with the help of area, productivity and season.

The two datasets were chosen for their richness, and are appropriate for training predictive models to recommend crops.

Data Preprocessing.

The data underwent some changes to make it ready for the recommender engine.

Loading the Data.

The Pandas library in Python was used to load the datasets, and the first inspection was done to see the datasets.

Handling Missing Values.

The isnull() function was used for finding missing values.

The dropna() method is used on the crop production dataset to drop missing values of production (896 Dropped).

Renaming and Cleaning Columns.

Changed names of columns for consistency in the data.

Standardized crop names were done (e.g., all names for moth were replaced by mothbeans).

We dropped some columns like State_Name and District_Name because they were not useful for our analysis.

Outlier Detection and Handling.

Histograms and box plots were used to identify outliers of continuous variables like N, P, K levels, temperature and rainfall.

As the extreme values were meaningful for agricultural scenarios, no transformation was performed.

Encoding Categorical Variables.

The crop recommendation dataset uses LabelEncoder to change the crop label into a numerical value which machine learning algorithms can use.

In the crop production data set, the categorical variables Season and Crop were converted to a one-hot encoded column using pd.get_dummies().

Correlation Analysis.

Using the Pearson correlation coefficient (PCC), numeric variables were studied for correlations. A heatmap was used to visualize the correlation matrix.

This stage helped to identify relationships between variables (e.g., production and area cultivated show a strong positive relationship).

Data Splitting.

Both datasets were divided into training and testing sets.

For the crop recommendation dataset, 80% was given for training and 20% for testing using the train_test_split method.

For the crop production dataset, the split ratio was 75 % training and 25 % testing in order to balance the larger dataset size.

Feature Scaling.

Min-Max scaling was applied to all continuous variables so that the same scaling range could be applied for better performance.

The dataset for the recommendation engine was removed of duplicates and standardized to make it ready for processing. The organized and cleaned data enabled us to train robust and efficient machine learning models as per agricultural requirements.

10.2. Complete Description of the Created/Downloaded Dataset

Dataset Overview.

The data given in this project is agricultural data which helps in giving suitable crop recommendations based on soil type, weather, productivity etc. Two datasets were used.

Crop Recommendation Dataset.

The dataset is from Kaggle- an open-source repository and implementation of this project used it directly.

Attributes: This data set has these process attributes.

The amount of nitrogen, phosphorus and potassium in the soil.

Thermometer: Heat measured in Celsius

Humidity: Percentage of relative humidity.

It denotes how acidic or alkaline the soil is.

The yearly rainfall, measured in millimeters.

The crop that is recommended based upon the input conditions.

The dataset features a total of 2,100 entries. Each entry combines the outcomes varied across the different input variables.

Classes: The dataset contains 22 (rice, maize, chickpea, etc.) distinct crop labels that can be taken in recommendation for the crops.

Crop Production Dataset.

The agricultural stats used here was available in public domain and obtained from Kaggle for use in this project.

Attributes.

The year a crop was grown is called the crop year.

The time of the year when the crop was grown, Kharif, Rabi, Summer, Winter.

Name of crop, or the crop.

The hectares allowed for the crop.

Amount produced in metric tonnes.

The dataset consists of 58,461 records related to different crops, seasons and measures of production over a few years.

Modeling User Preferences, Activities, and Goals

User Interests: The data aligns soil and environmental properties with productivity trends to offer tailored crop recommendations. It incorporates production and seasonal data to ensure practical and goal-oriented agricultural advice.

User Interactions: The "Crop Production Dataset" passively records user interactions through historical data on cultivated crops, their yields, and the conditions under which they thrived. These insights highlight successful crop cycles and provide data-driven recommendations for future planting.

User Intentions: By analyzing seasonal trends, soil nutrients, and weather patterns, the dataset supports systems designed to optimize crop productivity. Labeled recommendations act as practical guides, enabling users to achieve goals like maximizing land efficiency or selecting durable crops.

Dataset Descriptions:

Crop Recommendation Dataset:

No missing values are present.
Variables are organized and standardized for seamless model integration.
The categorical output (crop label) is encoded using LabelEncoder for compatibility with machine learning models.

EDA was conducted to identify trends, relationships, and anomalies within the dataset. Key findings include:

Variable Distribution

Nutrients (N, P, K):

Significant variations in nitrogen (N), phosphorus (P), and potassium (K) levels were identified using histograms and boxplots.
Crops such as rice and maize exhibited higher nitrogen levels, while blackgram and mungbean had lower phosphorus and potassium levels.
Outliers were detected, particularly in phosphorus and potassium distributions, prompting further analysis of their effects on crop suitability.
Temperature and Humidity:

Temperature ranged from 10°C to 40°C, with crops like maize and pomegranate thriving at higher temperatures.
Humidity varied from 20% to 100%, with crops like rice

Matched Source

Similarity 7%

Title:Investigating the effectiveness of one-class and binary ...

by JL Leevy · 2023 · Cited by 13 — Both datasets were divided into training and testing sets using an 80:20 ratio. To ensure balanced training for the OCC algorithms in the ...

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00825-1

Similarity 3%

Title:Effects of Different Amount of Biochar on Nitrogen, Phosphorus and ...

Jul 1, 2018 · Seven fertilization treatments are selected, combined with field experiments and indoor analysis methods to study the effects of biochar with nitrogen, phosphorus and potassium on soil nutrient and corn yield.

https://iopscience.iop.org/article/10.1088/1757-899X/394/2/022043