DATA ANALYSIS

Topic: CROP PRODUCTION Analysis

JULY 26

Authored by: Vishnu

Crop Production Prediction Project

Introduction:

The agriculture business domain, as a vital part of the overall supply chain, is expected to highly evolve in the upcoming years via the developments, which are taking place on the side of the Future Internet.

This paper presents a novel Business-to-Business collaboration platform from the agri-food sector perspective, which aims to facilitate the collaboration of numerous stakeholders belonging to associated business domains, in an effective and flexible manner.

This project aims to predict crop production based on the area of cultivation. We explore various machine learning models to determine the best approach for accurate predictions. The data used in this project includes the area of cultivation and the corresponding crop production.

This dataset provides a huge amount of information on crop production in India ranging from several years. Based on the Information the ultimate goal would be to predict crop production and find important insights highlighting key indicators and metrics that influence crop production.

To predict crop production, we would need to build a more sophisticated model using machine learning techniques. However, based on the insights we've gathered, here are some factors that would likely be important for prediction:

1. **State_Name**: The name of the state where the crop is produced.
2. **District_Name**: The name of the district within the state where the crop is produced.
3. **Crop_Year**: The year in which the crop production data was recorded.
4. **Season**: The season during which the crop was grown (e.g., Kharif, Rabi, Whole Year).
5. **Crop**: The type of crop that was grown (e.g., Rice, Wheat, Banana).
6. **Area**: The area of land (in hectares) used for growing the crop.
7. **Production**: The amount of crop produced (in tonnes).

DATA ANALYSIS

Data Overview:

- The dataset is labeled and contains information about crop production across various states and districts in India from 1997 to 2015.
- The **Area** column represents the area of land used for growing the crop (in hectares), and the **Production** column represents the amount of crop produced (in tonnes).
- The dataset has a wide range of values for both **Area** and **Production**, indicating diverse agricultural practices and crop yields across different regions and years.
- The dataset contains 246,091 entries with 7 columns: State_Name, District_Name, Crop_Year, Season, Crop, Area, and Production.
- The dataset covers 33 states, 646 districts, 6 seasons, and 124 different crops.
- Sugarcane appears to be the crop with the highest total production, followed by Rice and Wheat.
- Uttar Pradesh leads in total production, followed by Maharashtra and Punjab
- The production by season chart shows that Kharif (monsoon) season has the highest total production.
- Rabi (winter) season is the second most productive, followed by Whole Year crops.
- There are 124 unique crops in the dataset.
- The top 5 most frequent crops are Rice, Maize, Moong (Green Gram), Urad, and Sesamum.
- The dataset covers 33 states/union territories and 646 districts.
- Uttar Pradesh, Madhya Pradesh, Karnataka, Bihar, and Assam have the highest number of entries.
- The average production is about 582,503 units (likely in tonnes or quintals).

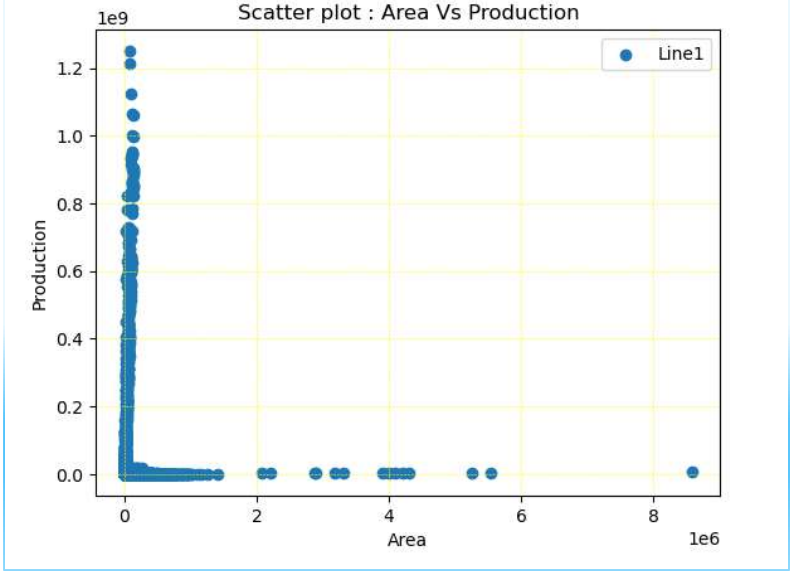
Data Preprocessing:

Let's start by preparing the data for the predictive model. We'll follow these steps:

1. Handle missing values: There are some missing values in the Production column (3,730 missing values).

2. Determining whether the data has non-linear relationships involves analysing the relationship between the features (independent variables) and the target variable (dependent variable).

3. To identify non-linear relationships in the data Visual Inspection using scatter plot:
    - Scatter Plots: Plotting the feature(s) against the target variable can help visualize the relationship. If the plot shows a curved or complex pattern rather than a straight line, it suggests a non-linear relationship.



The scatter plot shows the relationship between Area (x-axis) and Production (y-axis). From this plot, we can observe:

- There's a general upward trend, indicating a positive relationship between Area and Production.
- The relationship doesn't appear to be perfectly linear. We can see that as the Area increases, the spread of Production values also increases, suggesting some non-linearity.
- There are some outliers or extreme values, particularly for larger areas.

DATA ANALYSIS

Feature Selection:

Selecting features for continuous labeled data is an important step in many machine learning and data analysis tasks.
Correlation Coefficients: While Pearson's correlation coefficient measures linear relationships, Spearman's rank correlation can capture monotonic relationships, which might be non-linear.

Potential feature columns: Crop_Year, Area,State_Name, District_Name, Season, Crop.

Potential target column: Production

Categorical Data Columns: State_Name, District_Name, Season, and Crop are categorical and will need encoding

Label Encoding
This method assigns a unique integer to each category. It's useful for ordinal categorical variables where the categories have a specific order.

Example: If we have a Season column with values "Kharif", "Rabi", "Whole Year", label encoding will assign integers like:
"Kharif" -> 0
"Rabi" -> 1
"Whole Year" -> 2

Model Selection:

Given the nature of the data, a regression model is appropriate since the target variable (Production)data is continuous.

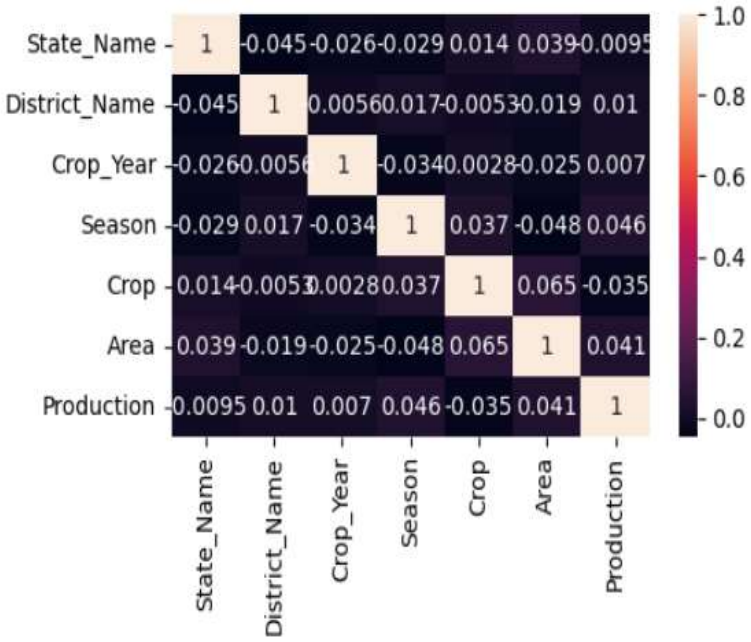
Linear Regression
Reason: Simple and interpretable model that works well if the relationship between the predictors and the target variable is linear.

Random Forest Regressor
Reason: Handles both linear and non-linear relationships well. It is robust to outliers and can handle missing values to some extent. Suitable for large datasets.

DATA ANALYSIS

Heat Map:

Visualize which predictors are most strongly correlated with the response variable.



High correlation between a predictor and the response variable indicates a potential linear relationship but as per the above heat map the correlation between the features and the target column are considerably weak, hence we are going to use the random regression model instead of the leaner regression model.

Model Training:

a. Split the data into training and testing sets.
b. Train a machine learning model (e.g., Random Forest).
c. determine the best model for crop production prediction, we would ideally:
d. Split the data into training, validation, and test sets.
e. Train multiple models on the training data.
f. Evaluate their performance on the validation set.
g. Fine-tune hyperparameters using techniques like cross-validation.
h. Finally, assess the best-performing model(s) on the test set.

DATA ANALYSIS

1. Data Loading and Preparation:

- The code starts by importing necessary libraries such as pandas, NumPy, scikit-learn, matplotlib, and seaborn.
- It loads the "Crop Production data.csv" file into a pandas Data Frame.
- The 'Area' column is selected as the feature, and 'Production' is set as the target variable.

2. Data Visualization:

- The code creates two visualizations of the 'Production' column:
  a. A boxplot to show the distribution and potential outliers.
  b. A histogram with a kernel density estimate (KDE) to show the overall distribution.

3. Outlier Removal:

- The code calculates the Interquartile Range (IQR) for the 'Production' column.
- It defines lower and upper bounds for outliers using the IQR method (Q1 - 1.5*IQR and Q3 + 1.5*IQR).
- Outliers are removed by filtering the Data-Frame to keep only the rows within these bounds.

4. Model Training and Evaluation:

- The filtered data is split into training and testing sets (70% train, 30% test).
- A Random Forest Regressor model is trained on the filtered data.
- The model makes predictions on the test set.
- The model's performance is evaluated using Mean Squared Error (MSE) and R-squared (R2) score.

5. Key Findings:

1. Data Cleaning: The original dataset had 246,091 rows, which was reduced to 199,971 rows after removing outliers.
2. Model Performance:

   1. **Mean Squared Error (MSE)**: The MSE value of approximately 27,147,621,935,643.957 indicates the average squared difference between the actual values and the predicted values..
   2. **Root Mean Squared Error (RMSE)**: The RMSE value of approximately 5,210,337.987 signifies the square root of the MSE, providing the measure of the standard deviation of the residuals or errors.
   3. **R-squared (R^2 score)**: The R-squared value of approximately 0.895 suggests that approximately 89.47% of the variance in the dependent variable can be explained by the independent variables in the model. A higher R-squared value closer to 1 indicates that the model fits the data well.