Weekly Progress Report

Pavan Balasaheb Kakde
Data science and Machine Learning
31/10/2025

Week Ending: 01

I. Overview:

This week, the primary focus was on understanding agricultural data and developing a prediction model for the project "Prediction of Agriculture Crop Production in India." Efforts were concentrated on data preparation, cleaning, merging multiple datasets, and building an initial regression model to predict crop production based on key agricultural factors such as cost, yield, and season.

II. Achievements:

1. Data Collection & Understanding

- Analyzed five raw datasets related to Indian crop production (2001–2014) obtained from data.gov.in.
- Identified key columns such as Crop, Variety, State, Cost, Yield, and Year for prediction.

2. Data Preprocessing

- Cleaned and merged datasets into a single file (final_crop_dataset.csv).
- Removed missing values, standardized units, and formatted data types.
- Saved processed data under the data/processed/ directory.

3. Model Building

- Trained a "Linear Regression" model to predict agricultural production.
- Calculated model accuracy using metrics like R^2, MAE, and RMSE.
- Visualized the relationship between actual vs predicted production values.

4. Result Analysis

- Found that Yield and Cost of Cultivation have the strongest correlation with Production.
- The model achieved an R^2 Score of ~0.88, showing strong predictive ability.

III. Challenges:

1. **Data Inconsistency**

- Faced issues due to non-matching column names and missing values across multiple CSV files.
- Resolved by manual column renaming and standardization.

2. **Model Limitations**

- Linear Regression did not capture complex non-linear patterns.
- Plan to test Random Forest Regression in the next phase for improved accuracy.

IV. Learning Resources:

1. **Machine Learning**

- Followed online tutorials on pandas, matplotlib, and scikit-learn for data cleaning and modeling.
- Referred to documentation from *scikit-learn.org* for regression and evaluation metrics.

1. **Data Visualization**

- Used Matplotlib and Seaborn for feature importance and correlation heatmaps.
- Analyzed model output using scatter plots and bar charts.

V. Next Week's Goals:

1. **Model Enhancement**

- Implement and test advanced models such as **Random Forest** and **XGBoost**.
- Compare their performance with Linear Regression.

2. **Result Documentation**

- Generate visualizations for model comparison and feature impact.
- Begin drafting the final report with observations, conclusion, and future scope.

VI. Additional Comments:

This week helped me gain practical experience in handling real agricultural data. Working on data cleaning, feature selection, and model building improved my understanding of how machine learning can be used to analyze and predict crop production effectively.