AI-Based Crop Yield Prediction

Glenn Noronha
West Texas A&M University
May 2025

Introduction

In this project, I created a machine learning model that uses real world weather, pesticide, and crop type data to predict crop yield. The main goal for this project was understanding which factors influence yield outcomes the most and to determine whether predictive modeling can help in making decisions for sustainable agriculture.

Instead of using already cleaned data sets, I searched for these data sets on the internet and ended up using raw data from reliable sources like Food and Agriculture Organization Statistics Database (FAOSTAT) and National Aeronautics and Space Administration (NASA) POWER. aligned the structure of the dataset to match those in a popular Kaggle crop yield dataset. This made it possible to test the model's performance under real world data constraints.

Dataset Overview

These are the datasets that were collected and used for data exploration:

Crop Yield Data: This data was sourced from FAOSTAT, and includes yield values (kh/ha) per crop and country from 2000 to 2023. There are many crops to choose from so I picked these six major crops: barley, maize (corn), potatoes, soya beans, wheat, and rice.

Weather Data: This data was sourced from NASA POWER. I had to retrieve the temperature (celsius) and precipitation (mm) data in separate files and then use data wrangling skills to merge them together.

Pesticide Usage Data: This data was also sourced from FAOSTAT, and includes data on the total amount of pesticides (kg/ha) used in each country.

Metadata: For model compatibility, country and crop names are one-hot encoded.

I took each dataset, cleaned it, and then merged them all together to create a cleaned crop data set. After preprocessing, the final dataset contained over 14,000 observations for 6 major crops and more than 150+ countries. This dataset contains features like average temperature in celsius, precipitation in millimeters, pesticide usage in kg/ha an, countrym and crop.

Data Preparation

A number of preprocessing steps were carried out in order to get the data ready to be modeled.

To enable easy merging of datasets, I had to convert the temperature and precipitation datasets from wide to long format. I then went through each dataset, removing any unnecessary columns and removing any rows with NA values. After that process was done, I merged all the data frames on the year column to have my final cleaned data frame.

One hot encoding was used to transform categorical variables like “Country” and “Item” into a ones and zeros so the model could read the data properly.

I used MinMaxScaler so the data values would be normalized between one and zero.

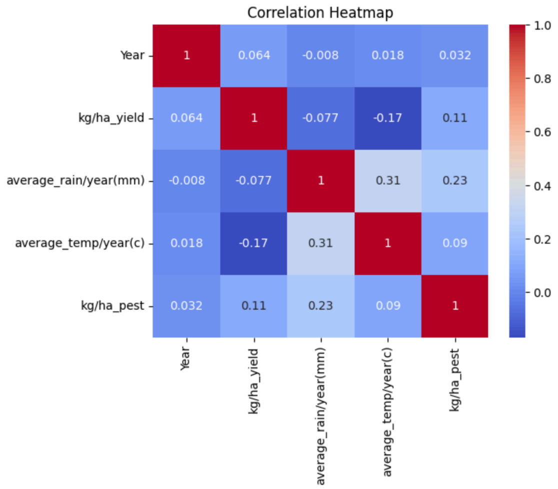
After all the data was done being cleaned and transformed, all the data sets were combined into a single data frame called “final_df”.

The target feature which was crop yield, was the y variable, and the modeling features mentioned earlier were inputted into a data frame named “features”.

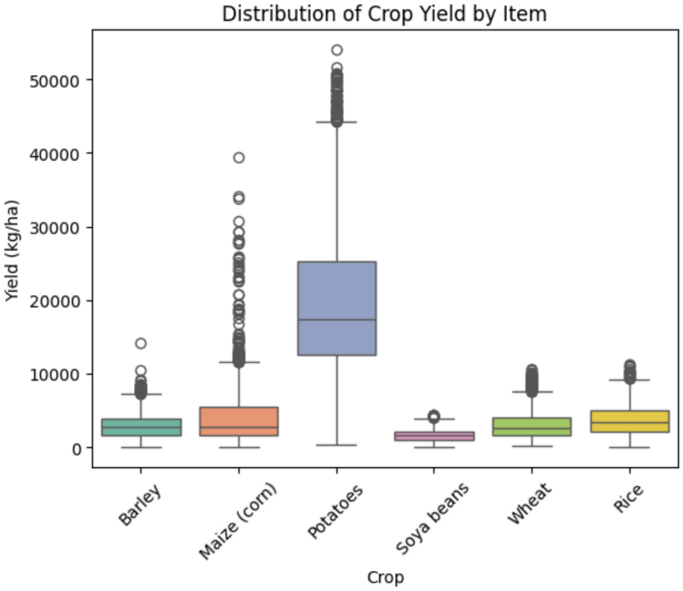
Exploratory Data Analysis

Various visualizations were used to better understand the structure and relationships within the dataset.

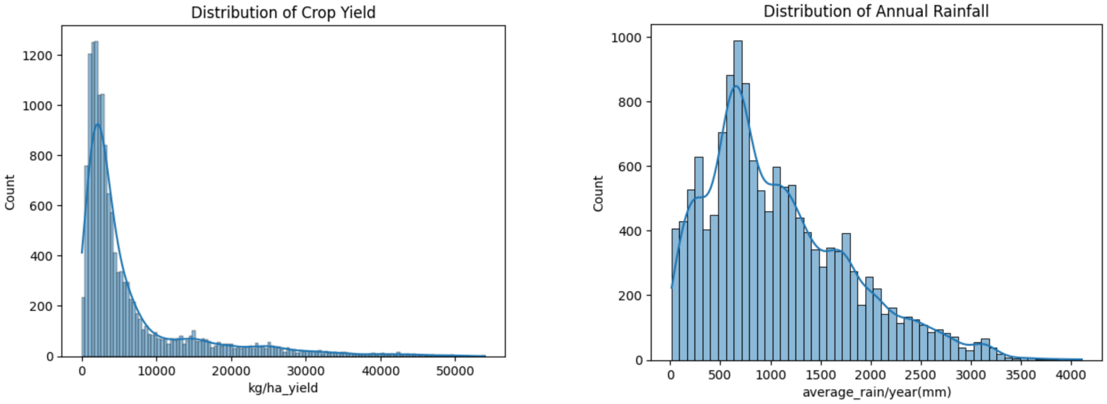
Heatmap of feature correlations showed pesticide usage and temperature to be positively correlated with yield.



Boxplots showed the range of yield for the different crops.



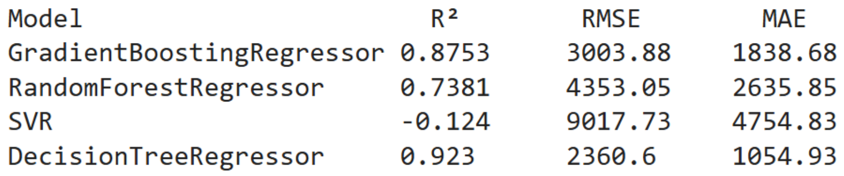
To see how features like yield and rainfall were distributed and skewed, distribution plots were used.



Feature importance plots (too big, can be viewed in notebook) derived from the Random Forest model showed that pesticides, temperature, and certain crop types (mainly potatoes) were some of the best predictors.

Model Training and Evaluation

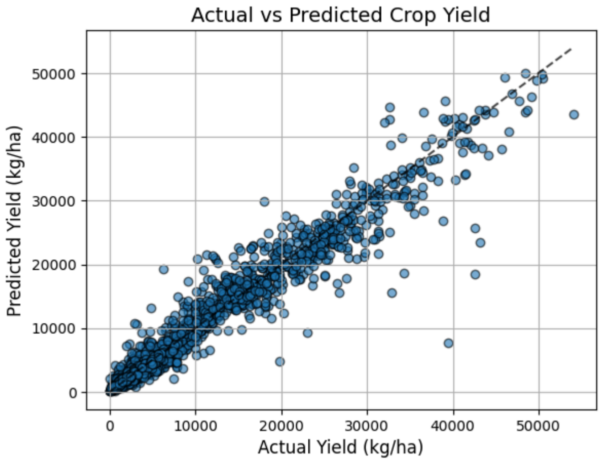
At first, I compared the performance of four different models (Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Regressor (SVR)) and used metrics like coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). When comparing models the Decision Tree and Gradient Boosting models appeared to perform extremely well, with strong R² scores and pretty low RMSE and MAE scores. However, when I evaluated the apparent top two models on each of the six crops, their performance was nowhere near as good with negative R² scores across the board. The one model that was able to maintain strong predictions across individual crops was the Random Forest Regressor. This model had consistent and strong R², RMSE, and MAE scores. This reminded me of the importance of not just evaluating models globally, but also within important subgroups in this case crop types.



```
Model                         R²         RMSE        MAE
GradientBoostingRegressor  0.8753     3003.88     1838.68
RandomForestRegressor      0.7381     4353.05     2635.85
SVR                        -0.124     9017.73     4754.83
DecisionTreeRegressor      0.923      2360.6      1054.93
```

Visualizations and Interpretations

I made some visualizations to better understand the Random Forest model’s behaviour. I did this by making a scatterplot of predicted vs actual data and it showed that the majority of the data points were closely grouped along the reference line. This indicates that the model’s prediction and the actual yield values were very similar. It is safe to say that for most situations, the model could produce accurate predictions.



Furthermore, the feature importance plot showcases that the most important factors influencing yield prediction were crop type, temperature and amount of pesticide usage.

These two graph visualizations were important to verify that the model was actually identifying significant relationships in the data.

Conclusion

This project showed how to use real world, multi source data to create an interpretable and accurate machine learning model for predicting crop yields. I was able to combine weather, pesticide, and yield data into a single data frame through thorough preprocessing steps. When various models were tested, the Random Forest Regressor model proved to be the most dependable and maintained high performance even when predicting crop by crop yield.

The most important lesson of this project is that in order to accurately predict real world data, the data must be carefully cleaned, processed and scaled. It is also apparent that not all crops would benefit from a single model, which shows the value of evaluating performance across sub groups.

References

FAOSTAT (Food and Agriculture Organization):
Food and Agriculture Organization of the United Nations. *FAOSTAT*. https://www.fao.org/faostat

World Bank Climate Data:
World Bank. *Climate Knowledge Portal*. https://climateknowledgeportal.worldbank.org/