Choose Your Own - Smart Agriculture Project

Capstone Project – Choose Your Own Dataset
HarvardX PH125.9x – Data Science Capstone
Felipe Muniz
June 2025

Introduction

This project applies machine learning to a smart agriculture dataset. The objective is to recommend suitable crops based on environmental and soil characteristics. The dataset includes 2200 observations with the following variables: Nitrogen (N), Phosphorous (P), Potassium (K), temperature, humidity, pH, and rainfall. The target variable is the crop type. Although the dataset includes several agronomic variables, pH and rainfall were selected for visual exploration due to their relative simplicity to visualize and their ability to reveal crop-specific environmental niches. Nutrient concentrations (N, P, K) are critical features but do not yield as visually distinct clusters when explored in two-dimensional plots without advanced transformation or domain-specific calibration. Modeling methods like Random Forest are well-suited to incorporate these hidden interactions.

Methods

The data was imported from Kaggle[1] repository and selected to answwer the demands of the instructions for the current assignment. The comma-separated file (CSV) converted to a classification-friendly format. No missing values were found, and no additional cleaning was required.

An **80/20 train-test split** was implemented using `createDataPartition()` from the `caret` package. This proportion balances the availability of data for training with a fair estimate of model performance.

Two models were trained: - **Multinomial Logistic Regression**, a linear baseline model for multiclass classification. - **Random Forest**, an ensemble decision tree algorithm chosen for its ability to model non-linear relationships and interactions.

Below are two exploratory visualizations to understand crop distributions under soil and climate conditions:

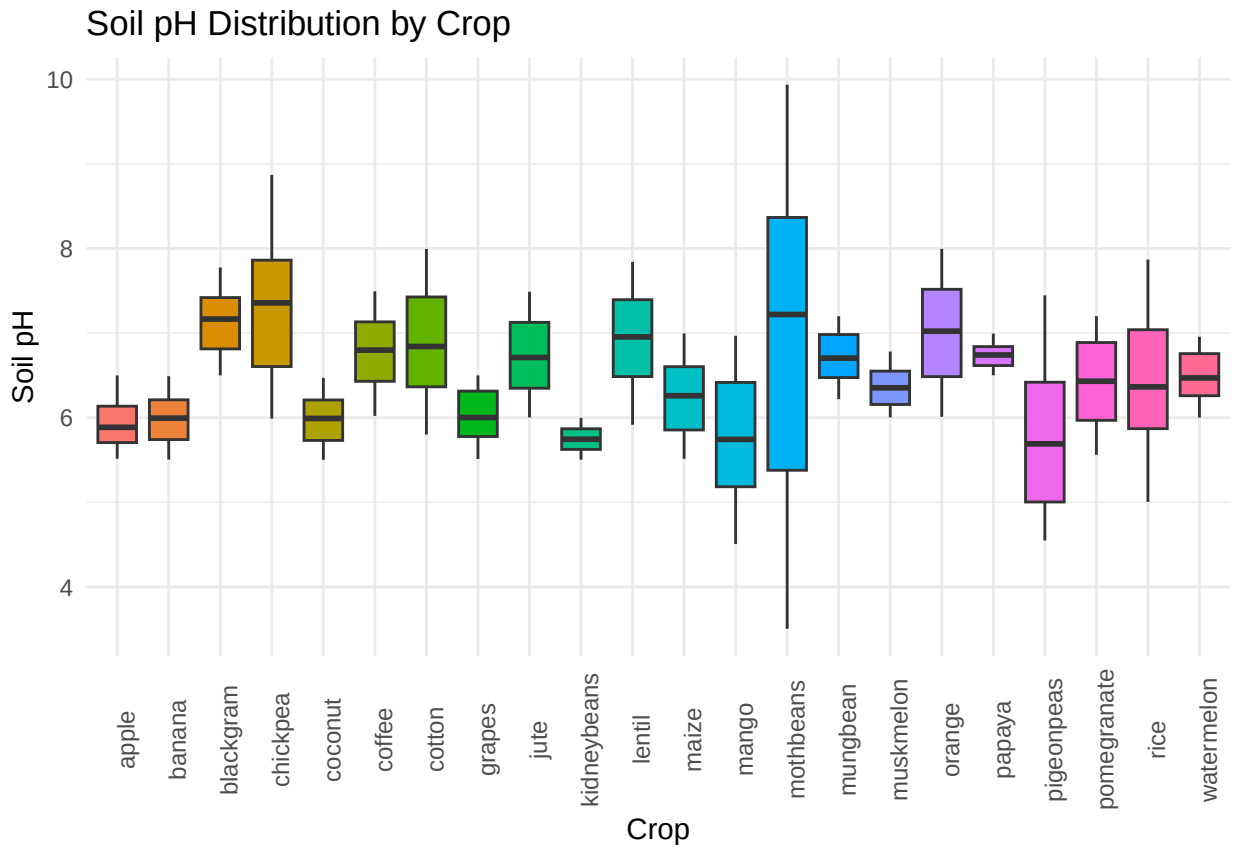
Boxplots of pH and Rainfall by Crop Type

The visualization uses boxplots to show the full distribution of pH and rainfall for each crop. Boxplots were chosen due to their practicability to compare several crops in terms of rainfall and soil pH profile requirements.

Plot Soil pH

This boxplot suggests that almost all legumes, together with rice and cotton thrives across the widest range of pH, with mothbean particularly able to thrive in both acidic and alkaline soils. In fact, legumes are the crops able to thrive in alkaline soil profiles, except for kidney beans. Most crops thrive between pH 6 and 7.5, suggesting these crops tend to develop in acid soil profiles. Interestingly enough, kidney beans together with papaya, showed the narrower tolerances to pH conditions.

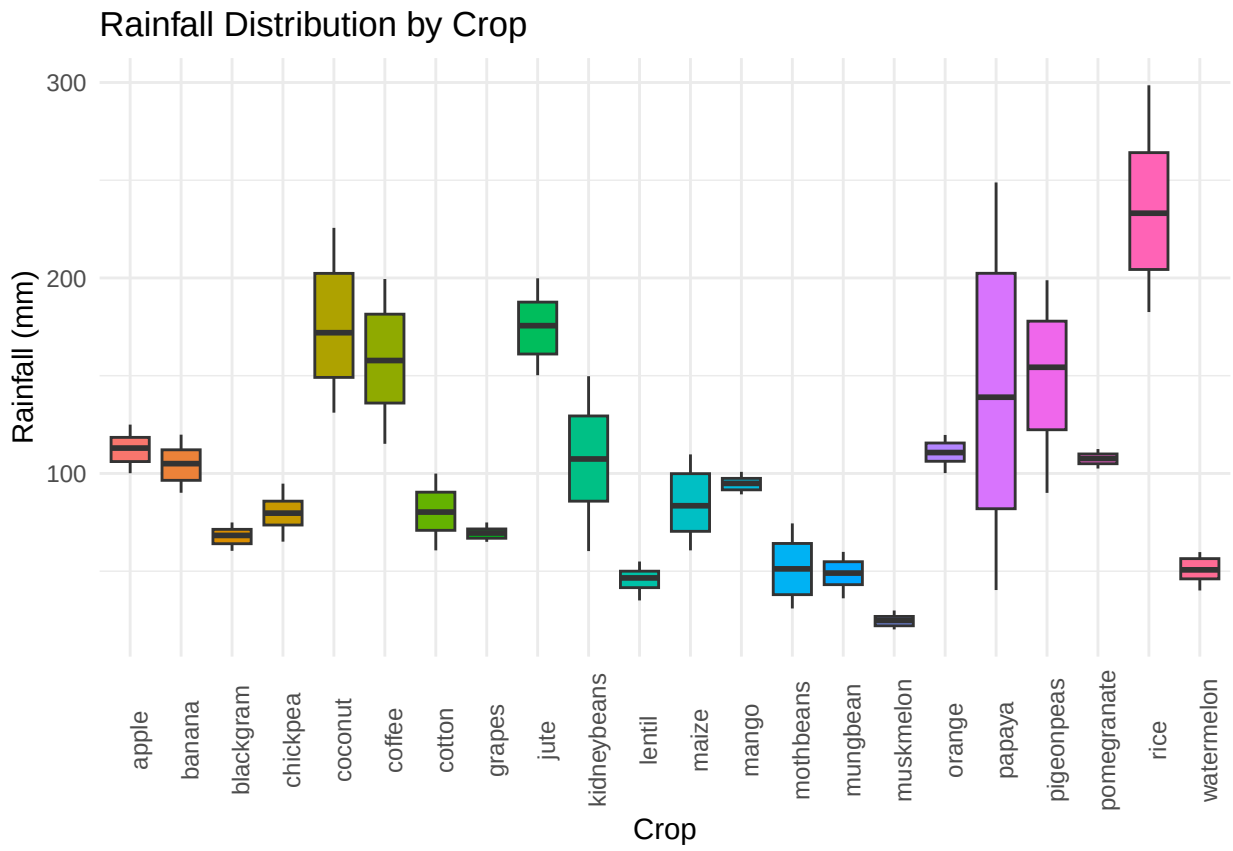
```
# Boxplot of pH by crop
df <- read.csv("Crop_recommendation.csv")
df$label <- as.factor(df$label)
ggplot(df, aes(x = label, y = ph, fill = label)) +
  geom_boxplot(outlier.shape = 16, outlier.colour = "red", outlier.size = 2) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Soil pH Distribution by Crop", x = "Crop", y = "Soil pH") +
  guides(fill = "none")
```



Plot Rainfall

The rainfall distribution suggests that most crops develop between 50mm and 200mm rainfall precipitation values, with exception to rice that is positioned between 200mm and somewhere above 250mm. Pomegranate, watermelon, muskmelon, grapes, black gram, orange are crops that demand a particular rainfall value. Most crops have their associated rainfall values between 50mm and 150mm.

```r
# Boxplot of rainfall by crop
ggplot(df, aes(x = label, y = rainfall, fill = label)) +
  geom_boxplot(outlier.shape = 16, outlier.colour = "red", outlier.size = 2) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Rainfall Distribution by Crop", x = "Crop", y = "Rainfall (mm)") +
  guides(fill = "none")
```



Results

Model training and testing were performed using the following code:

```r
# Set seed for reproducibility
set.seed(123)

# Create a stratified 80/20 split for training and test sets
index <- createDataPartition(df$label, p = 0.8, list = FALSE)
train <- df[index, ]
test <- df[-index, ]

# Train a multinomial logistic regression model
log_model <- multinom(label ~ ., data = train, trace = FALSE)
```

```r
# Predict labels on the test set using the logistic model
log_pred <- predict(log_model, test)

# Calculate accuracy for logistic regression
log_acc <- mean(log_pred == test$label)

# Train a random forest model
rf_model <- randomForest(label ~ ., data = train)

# Predict labels on the test set using the random forest
rf_pred <- predict(rf_model, test)

# Calculate accuracy for random forest
rf_acc <- mean(rf_pred == test$label)

# Print both accuracy scores
log_acc
```

```
## [1] 0.85
```

```r
rf_acc
```

```
## [1] 0.9931818
```

The random forest model achieved an accuracy of over 99%, significantly outperforming logistic regression, which reached 85%. This result confirms the suitability of tree-based ensemble methods for agricultural data modeling.

Conclusion

A functional crop recommendation model was developed using basic environmental and soil data. The modeling process demonstrated how even minimal input features can drive high classification performance when using appropriate algorithms.

While the results are promising, limitations include the lack of geospatial and seasonal data. Future work could include the integration of time-series weather data, satellite imagery, or localized soil profiles to improve model accuracy and adaptability.

References

1. Kumari, C. (2022). *Smart Agricultural Production Optimizing Engine* [Dataset]. Kaggle. https://www.kaggle.com/datasets/chitrakumari25/smart-agricultural-production-optimizing-engine/data

2. Irizarry, R. A. (2023). *Introduction to Data Science: Statistics and Prediction Algorithms Through Case Studies* (2nd ed.). Leanpub.

3. Irizarry, R. A. (n.d.). *Introduction to Data Science: Data Wrangling and Visualization with R* [Online course]. HarvardX PH125.1x – PH125.2x. edX.