Content Page

Introduction

Agriculture is among the oldest and most useful branches of the human civilization as it provides them with a source of food, raw materials, and economic stability. As big data and machine learning develop, there are new opportunities to analyze the environment and make more reasonable decisions regarding crop production. These innovations are critical in food security in nations such as India, where agriculture contributes to the livelihood of millions of population.

The data that this project is focused on is an agricultural dataset which consists of both environmental variables and labels of crops. The dataset was chosen due to its potential effect in the real-world: the research of the influence of soil and climatic factors on crops suitability may allow optimizing the agricultural activity, increase yields, and aid in policy formulation.

The aims of this project were two:

- To conduct exploratory data analysis (EDA) in order to determine patterns, distributions, and groupings in the dataset.
- To construct a regression model to analyze whether rainfall can be predicted by using the independent variables temperature, humidity and soil pH.

Although EDA proved that crops have very strong relationship with climate and soil conditions, the regression model showed that there are some limitations regarding the prediction of rainfall. This result supports the significance of viewing more sophisticated or alternative methods, e.g., non-linear models and classification methods, in future agricultural analytics.

Data

Dataset Source and Structure:

- The data was selected by going through Indian government repositories and cleaned by Kunsh Bhatia.
- It has been made ready to do predictive modelling, and it is aimed to aid in the research of agricultural advice.

Files included:

- Crop_recommendation.csv - raw data
- Crop Data.xlsx.csv - processed and cleaned data.

Dimensions of the dataset: 2201 records and 6 features.

Field	Description	Data Type	Variable Type
Temperature	Average temperature (°C)	Float	Continuous
Humidity	Average humidity (%)	Float	Continuous
pH	Soil pH value	Float	Continuous
Rainfall	Average rainfall (mm)	Float	Continuous
Label	Crop name (e.g., rice, wheat)	String	Categorical
Label_Num	Encoded crop label	Integer	Categorical

Both classification tasks (e.g., crop recommendation) and regression tasks (e.g., rainfall prediction) can be performed by this structure.

Data Cleaning Steps:

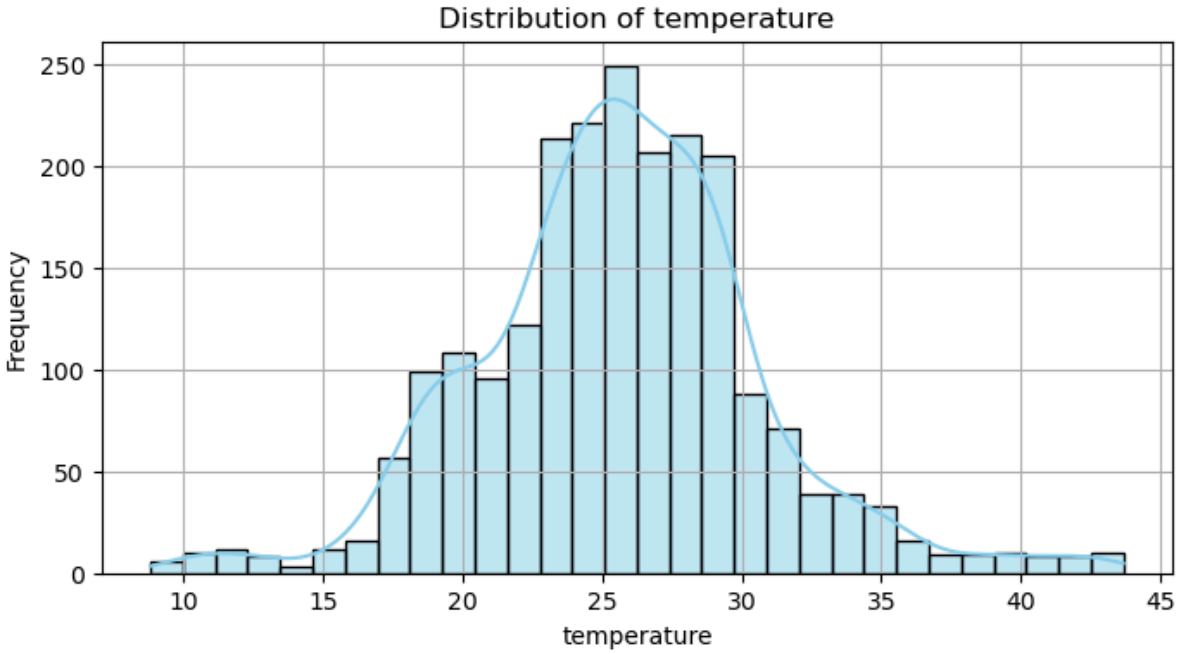
- This information had been preprocessed but additional procedures were carried out.
- Duplicate Check: There is none, but deleted to be on the safe side.
- Missing Values: none.
- Data Type Check: All values that are floats stored correctly.
- Label Normalization: Name names normalized (lower case, no spaces).
- Label Validation: Checked on 22 known types of crops (e.g. rice, maize, banana, mango).

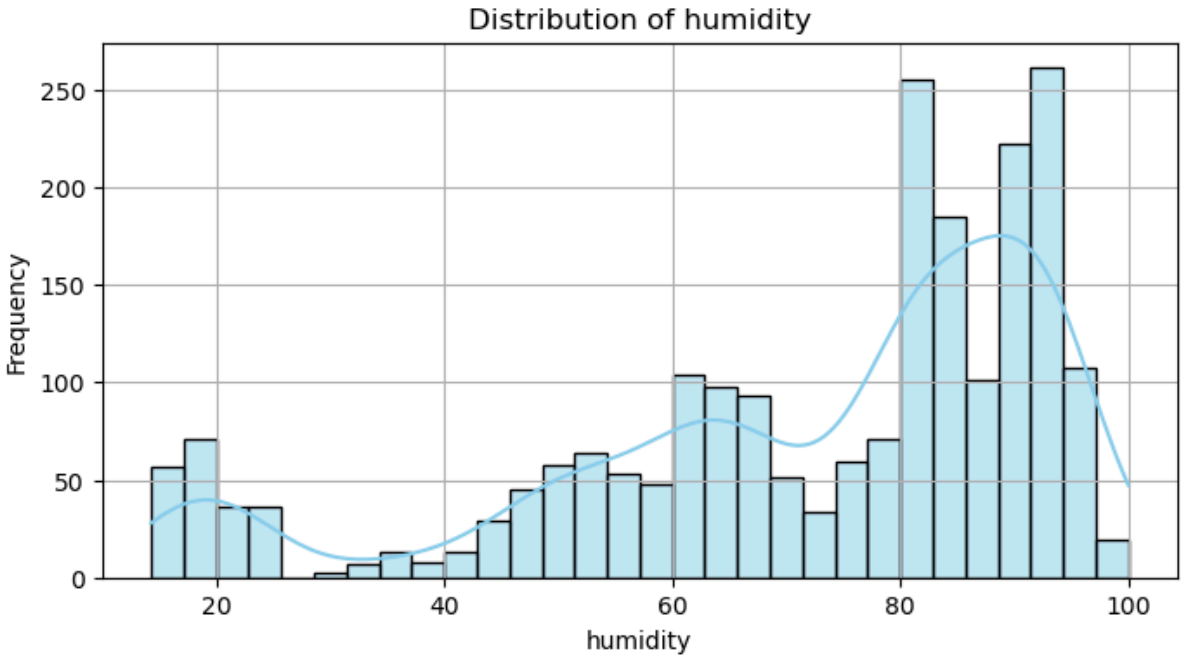
Sample of cleaned data:

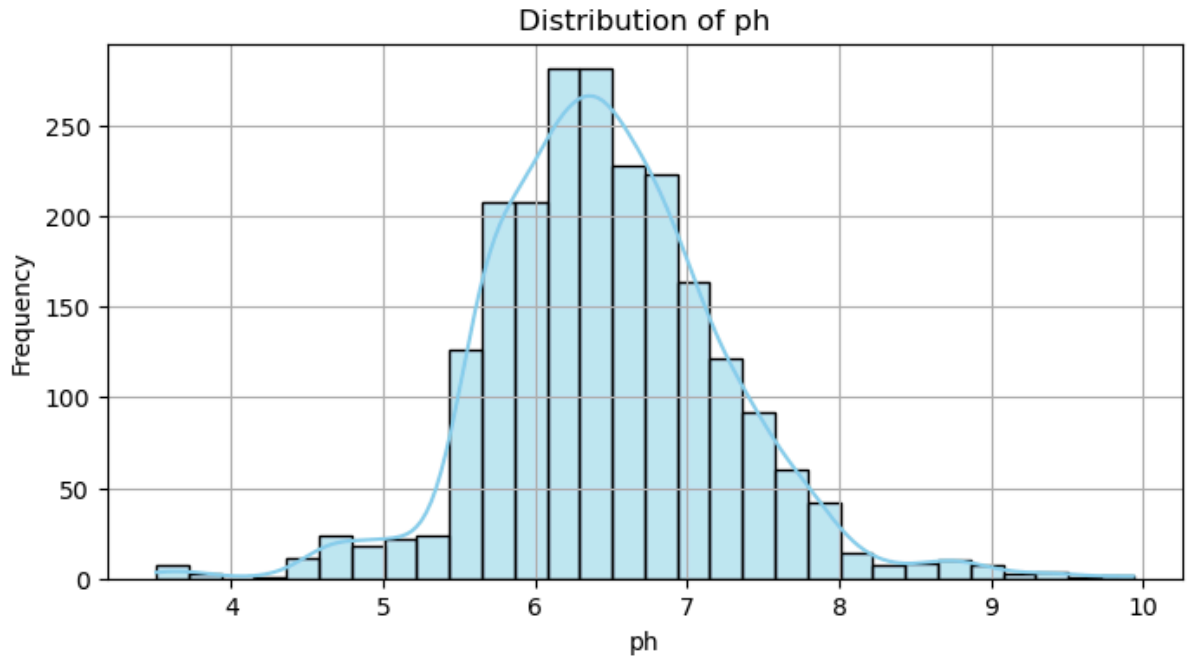
temperature	humidity	ph	rainfall	label	Label_Num
20.87974371	82.00274423	6.502985292	202.9355362	rice	0
21.77046169	80.31964408	7.038096361	226.6555374	rice	0

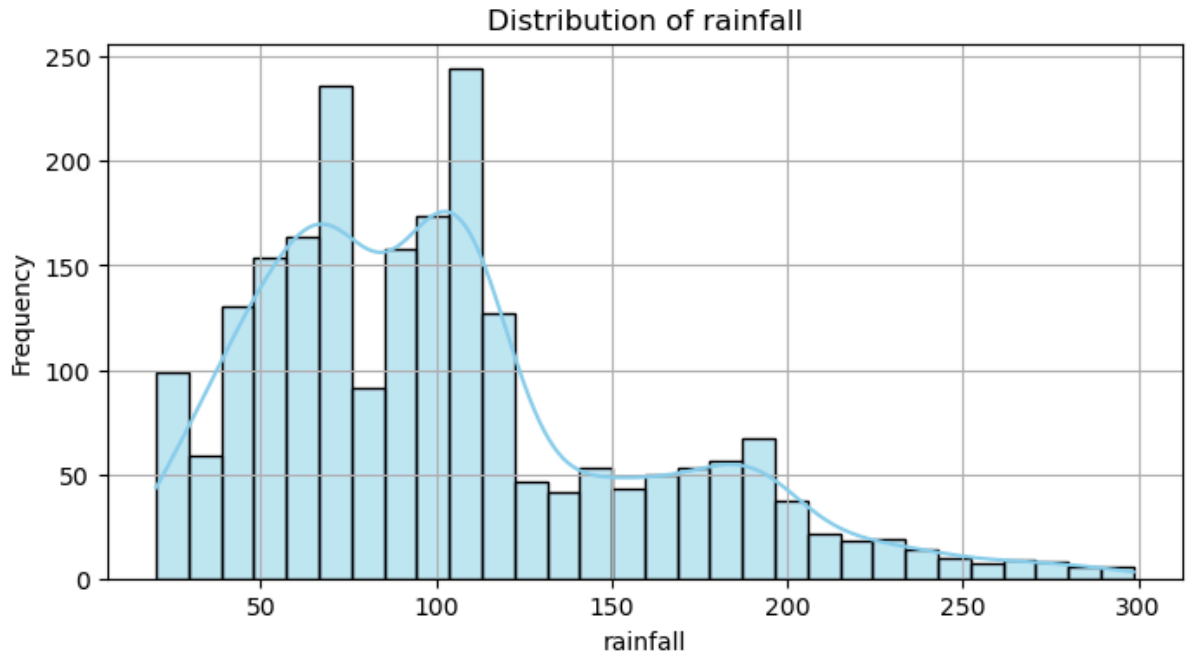
Exploratory Insights from Distributions

- Temperature: Distribution of about 25-28 o C. Outliers observed at ~10°C and ~40°C
- Humidity: Multi-modal distribution; peaks at ~20–30%, 40–60%, and 80–100%. This indicates climatic plasticity amongst crops.
- pH: In the 5.5-7 range (neutral to a little acidic soil). Few crops tolerate <4.5 or >9 pH
- Rainfall: Right-skewed; majority of values between 50–150 mm. Some crops extend to 200–300 mm









Methods

Two techniques were used:

Exploratory Data Analysis (EDA)

- Boxplot was utilized to contrast the suitability of crops with the humidity, temperature and pH levels.
- Frequency distribution was determined using histograms.
- The variables were evaluated on the basis of heatmaps and pairplots to determine correlations and potential multicollinearity.

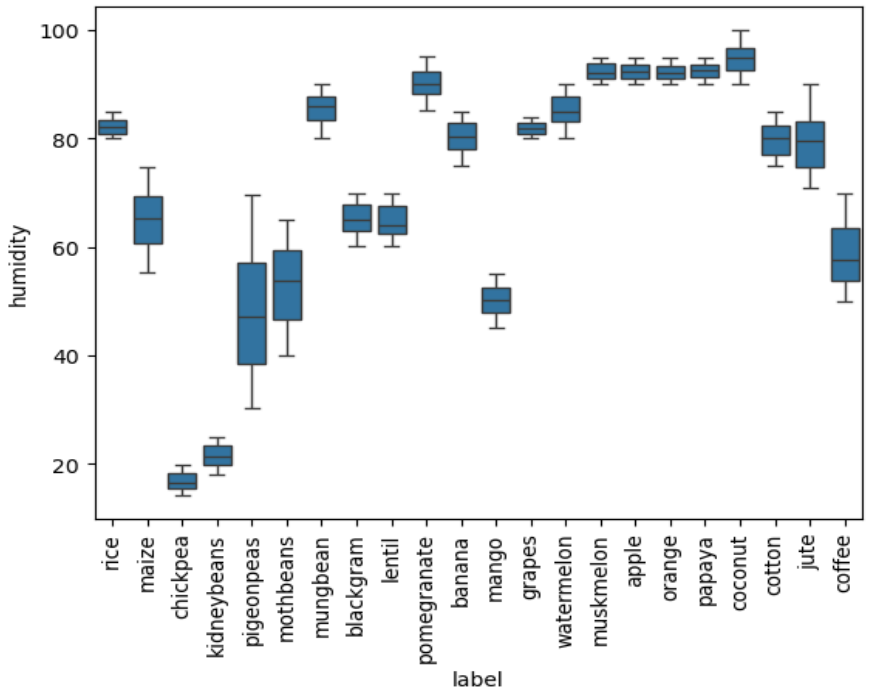
Regression Modeling:

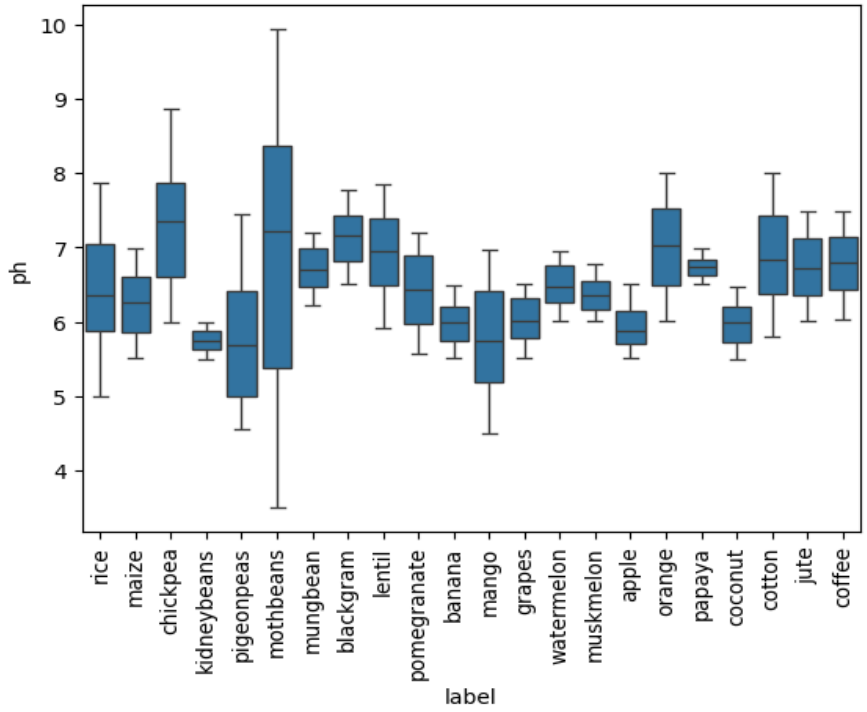
- Linear regression model was developed using rainfall as the dependent variable and the independent variables temperature, humidity and pH.
- Evaluation Metrics:
    - R 2 to determine explained variance.
    - RMSE (root mean square error) to measure accuracy of prediction.
- The performance was evaluated visually by using scatterplots of predicted vs. actual rainfall.

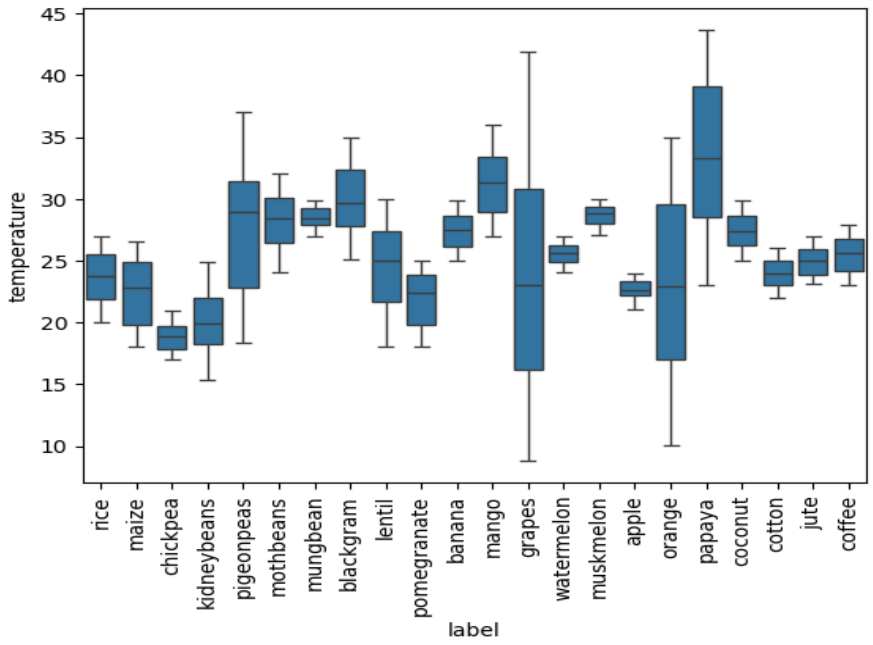
Analysis

Boxplot Analysis

- Humidity vs. Crop:
    - Rice: thrives in very high humidity (80-90%).
    - Chickpea & Lentil: low humidity preference (20-40%).
    - Coconut/Maize: constant level of humidity with minimal fluctuation.
- pH vs. Crop:
    - Most crops like 5-7 pH.
    - Chickpea shows wide adaptability.
    - Coffee/jute restricted to a limited range of pH.
- Temperature vs. Crop:
    - Rice, maize, sugarcane - 28-35degC (warm season crops).
    - Wheat, barley, potato - 15-22degC (cool season crops).
    - Cotton spans a wider range




ph
label
rice
maize
chickpea
kidneybeans
pigeonpeas
mothbeans
mungbean
blackgram
lentil
pomegranate
banana
mango
grapes
watermelon
muskmelon
apple
orange
papaya
coconut
cotton
jute
coffee

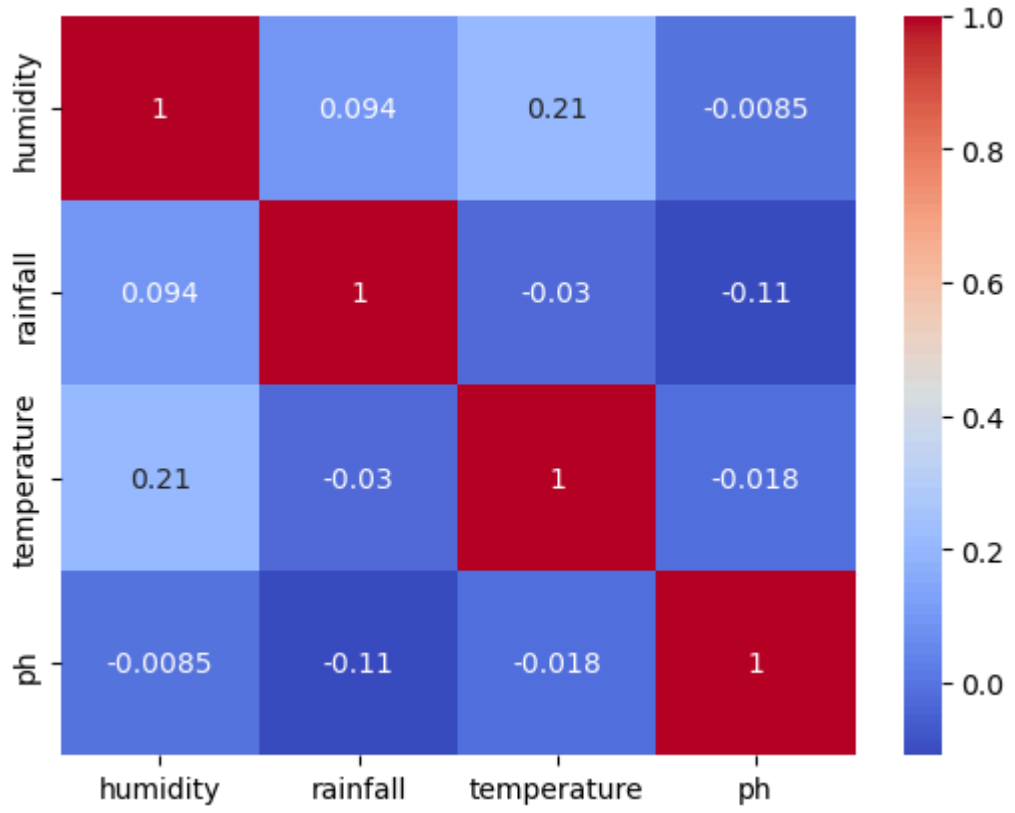

temperature
label
rice
maize
chickpea
kidneybeans
pigeonpeas
mothbeans
mungbean
blackgram
lentil
pomegranate
banana
mango
grapes
watermelon
muskmelon
apple
orange
papaya
coconut
cotton
jute
coffee

Heatmap and Correlation

The relationship was weak as shown in the correlation matrix.

- Temp vs. Humidity: 0.21 (weak positive)
- Rainfall vs. pH: -0.11 (weak negative)
- Rainfall vs. Temp: -0.03
- Temp vs. pH: -0.018

Interpretation: No strong predictor of rainfall exists among these variables. This independence implies that they can still be applied to classification tasks without issue of multicollinearity.



Regression Model Results:

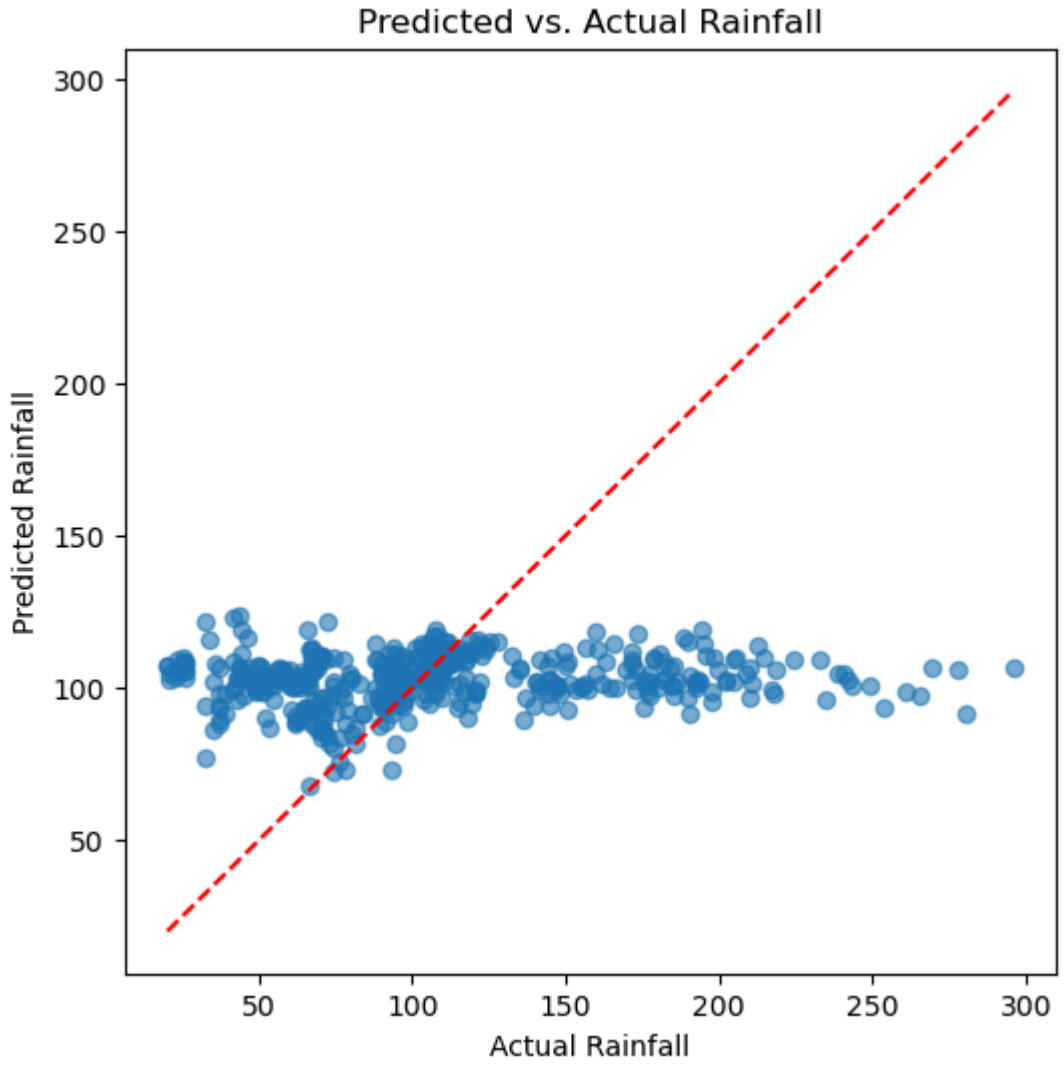
Intercept: 159.73749749637517
Slopes: [-0.52316338 0.214176 -9.11378845]
R^2 score: 0.009138453861105233
RMSE: 54.615447047587566

- Intercept: ~160 mm
- Slopes: Temperature (-0.52), Humidity (+0.21), pH (-9.11)
- R2: 0.009 (accounts less than 1 percent of the rainfall variance).
- RMSE: ~55 mm (large prediction errors).

Scatterplot Analysis:

- The forecasts concentrated on predictions of 100-120 mm irrespective of real rainfall.
- Reduced rain level (>150 mm).
- Excessive low precipitation (less than 100 mm).

Interpretation: The Regression model showed that rainfall cannot be predicted (linearly) using the chosen environmental features.



Conclusion

The project was able to explore the agricultural data and reach the following:

- Patterns of Crop Suitability:
    - Easy to see groups (e.g. rice and sugarcane require high humidity/temperature, wheat and barley can be grown in cooler damp climate).
    - Crops differ greatly in their environmental conditions such as temperature, humidity and pH.
- Regression Model Limitations:
    - Prediction of rainfall was ineffective (R 2 = 0.009, RMSE= 55 mm).
    - Rainfall variability could not be explained by independent variables, which proved the hypothesis according to which linear regression is not appropriate to perform this task.
- Future Prospects:
    - Decision trees (random forests, etc.) classification models would be more applicable towards crop recommendations.
    - Rainfall prediction with non-linear models may be tested, and some extra characteristics of soil type, geographic location, or seasonal trends would be considered.

To conclude, the regression approach was unsuccessful, but the overall project still offers a great insight into the field of agricultural analytics and suggests more promising future trends in machine learning as applied to agriculture.