Predict Crop Yield based on Environmental and Agricultural Factors

1st Arpit Dilip Sharma
Computer Science and Engineering
SUNY Buffalo
Buffalo, NY, USA
arpitdil@buffalo.edu

2nd Darshan Sunil Sonawane
Computer Science and Engineering
SUNY BUffalo
Buffalo, NY, USA
dsonawan@buffalo.edu

***Abstract*—Agriculture plays significant part in lot of countries GDP. This project aims to predict crop yield, measured in tons per hectare based on environmental and agricultural data which includes variables like rainfall, temperature, etc. To get results extensive data cleaning steps are performed to have best data quality by performing operations on uncleaned data. This will help in optimizing agricultural productivity and decision-making for farmers.**

I. Introduction

Crop yield is important for many reasons including Food security, finding solutions to reduce food waste, environmental impact, market prices. Therefore, calculating the crops yielded in tons per hectare plays a significant role. In our project we aim to calculate this with the help of some key features including region, rainfall, temperature and many more. The initiative intends to help farmers and agribusinesses make educated decisions that improve productivity and sustainability by developing a strong predictive model.

II. Problem Statement

The objective of this project is to estimate the crop yield in tons per hectare based on different environmental inputs such as rainfall, temperature, soil type and agricultural inputs like irrigation, fertilizer usage which are some agricultural inputs. Key question is: How can crop yield be predicted based on environmental and agricultural factors?

III. Potential and Contribution

The purpose of this project is to construct an accurate crop yield prediction model in order to make a contribution to the field of agricultural data analysis. This input is essential for a number of reasons:

1) Data-Driven Decision Making: It presents a method for yield prediction based on data, which lowers uncertainty in farming decisions and encourages more strategic, educated choices.
2) Resource Optimization: The model's ability to predict yields can be used to optimize vital agricultural inputs like fertilizer and irrigation, resulting in more sustainable farming methods and more effective use of resources.
3) Generalizability: Should the model's development be successful, it might be applied to a wide range of crops and geographical areas, which would increase its influence on global agricultural productivity.
4) Data Integrity: The project highlights the significance of preparing data to overcome problems like missing values and inconsistent datasets, ensuring the accuracy and reliability of the predictive model.

By creating a robust and adaptable yield prediction model, this project endeavors to provide a valuable tool for enhancing agricultural productivity and fostering sustainable farming practices.

IV. Cleaning steps performed

A. Removing Duplicates

First step is to identify and removing duplicate rows in the dataset to ensure data consistency and prevent biases in predictive modeling.

B. Replacing missing values

Another cleaning step which performed is imputation techniques on numeric columns to remove the null values and replace it with the mean value of that column.

C. Remove white-spaces

This step performs operation on string type column to remove leading and trailing white-spaces. Whitespaces can lead to inconsistency in future for data processing.

D. Standardize boolean values

Another step for data cleaning is to standardizing the boolean value columns for this project it is Fertilizer used and Irrigation Used.

E. Standardize string values

This step is performed to bring consistency in all string columns to make it upper case, to avoid any typo error in the data.

F. Categorical data conversion

Converting string data to categorical data helps to improve memory usage and significantly improve analytical efficiency. This is done using pandas library's command astype() command.

G. Numerical data consistency

Rainfall mm column have a float type value which contained lot of precision after decimal place. This step is performed to bring consistency in column and make decimal precision to 2.

H. Remove outliers

To ensure data quality and avoid any influence of any extreme values present in dataset. The interquartile range method was used to identify outliers from numeric columns.

I. On-hot-encoding

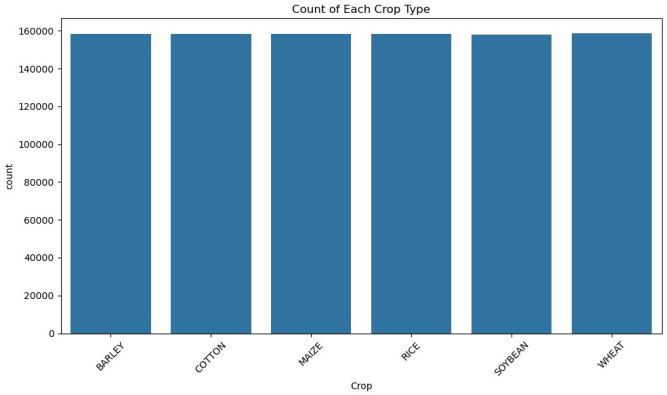
To ensure use of machine learning algorithms that require numerical input, categorical features were transformed into a numerical representation using one-hot encoding. The pd.get_dummies() function was used to create binary (0/1) columns for each unique value within the categorical columns Region, Soil_Type, Crop, Weather_Condition, Irrigation Used, and Fertilizer_Used.

J. Normalization

Final cleaning step facilitate meaningful comparison between features, the Rainfall_mm and Temperature_Celsius columns were normalized. Min-Max scaling was applied using the MinMaxScaler from the sklearn.preprocessing library, effectively transforming the values in these columns to a common scale between 0 and 1.

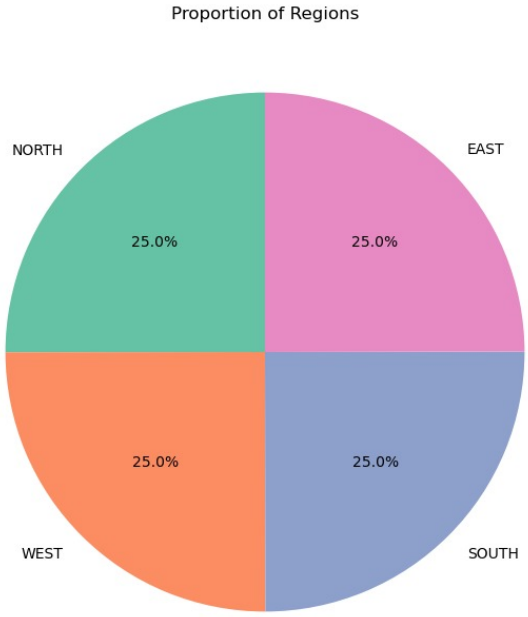
V. Exploratory Data Analysis

A. Count of each crop type



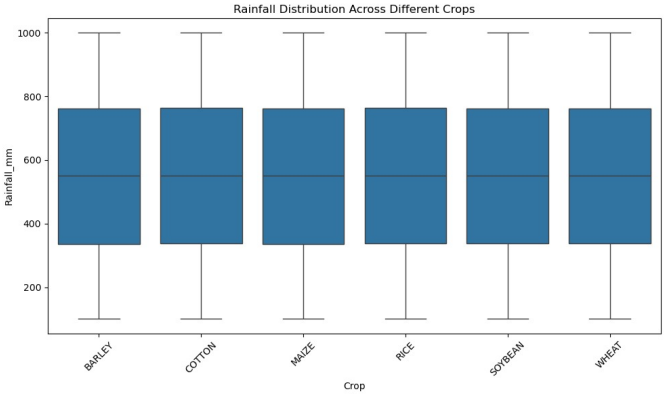
The count plot titled ”Count of Each Crop Type” shows a nearly uniform distribution of different crops in the dataset. Each crop type appears approximately 160,000 times.

B. Proportion of regions



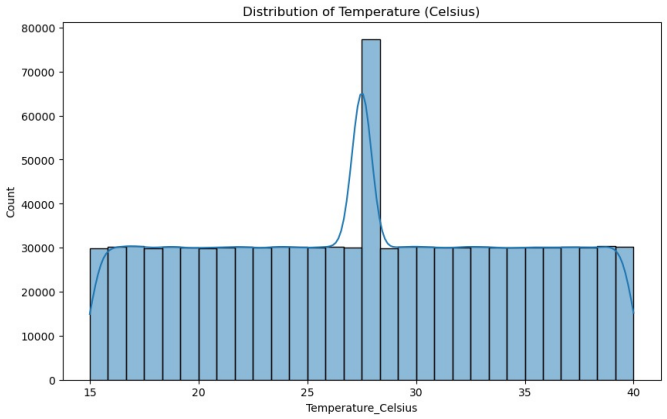
The pie chart titled ”Proportion of Regions” reveals an equal distribution of data across four regions (North, East, South, and West), with each region accounting for 25 of the dataset.

C. Distribution of rainfall across different crops



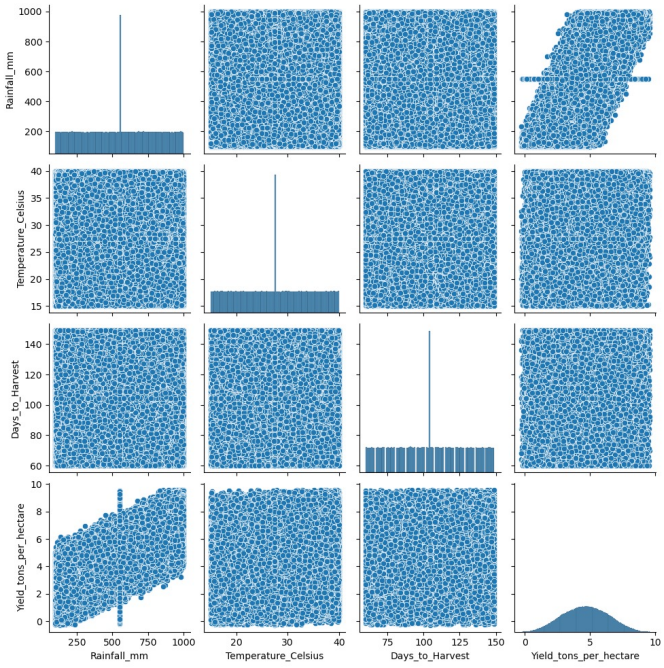
The Box plot titled ”Rainfall Distribution Across Different Crops” shows a similar distribution of rainfall across all crop types (Barley, Cotton, Maize, Rice, Soybean, and Wheat).

D. Distribution of Temperature



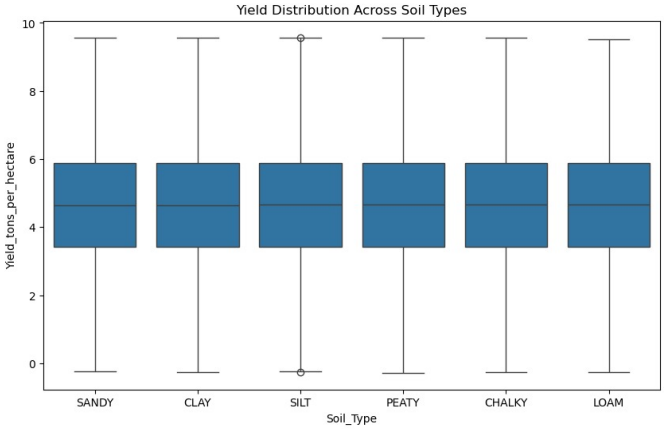
The Histogram titled "Distribution of Temperature (Celsius)" shows a multimodal distribution with a prominent peak around 28°C and smaller peaks around 17°C and 38°C.

E. Relationship among continuous variables



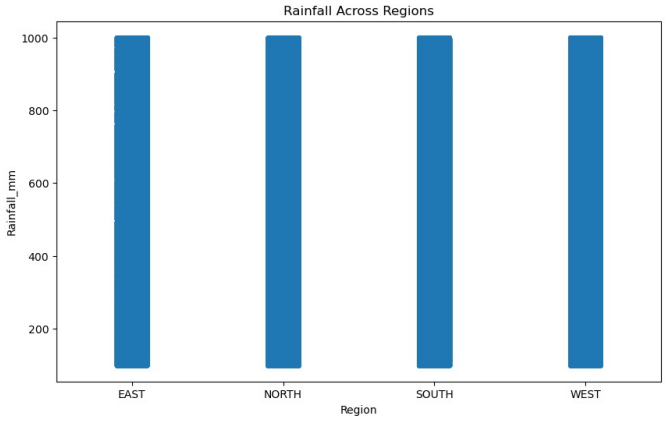
The pair plot shows the relationships between Rainfall_mm, Temperature_Celsius, Days_to_Harvest, and Yield_tons_per_hectare. A strong positive correlation is evident between Days_to_Harvest and Yield_tons_per_hectare

F. Yield Distribution across soil types



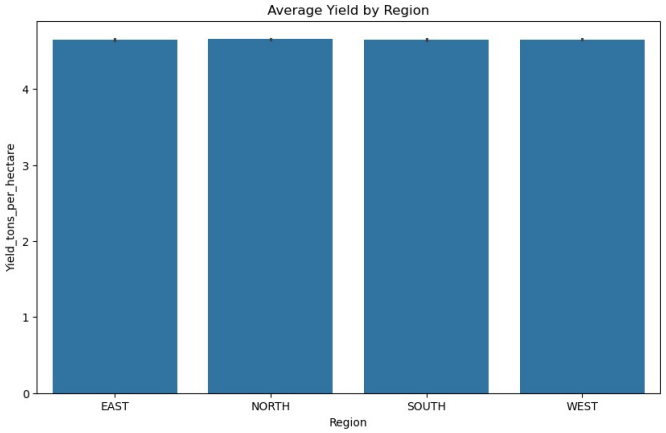
Box plot titled "Yield Distribution Across Soil Types" reveals that different soil types exhibit similar yield distributions, with medians hovering around 5 tons per hectare.

G. Rainfall across regions



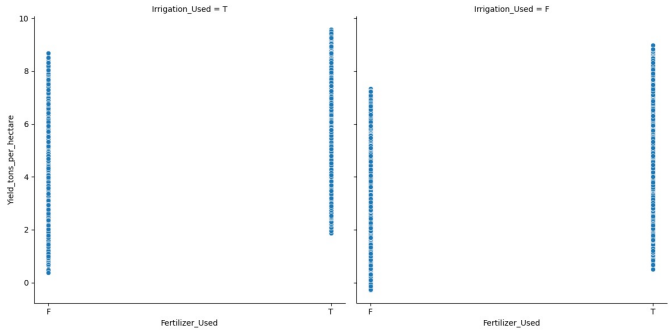
Strip plot "Rainfall Across Regions" indicates a similar distribution of rainfall across all four regions (East, North, South, and West), with most data points concentrated between 200 mm and 1000 mm.

H. Average Yield by region



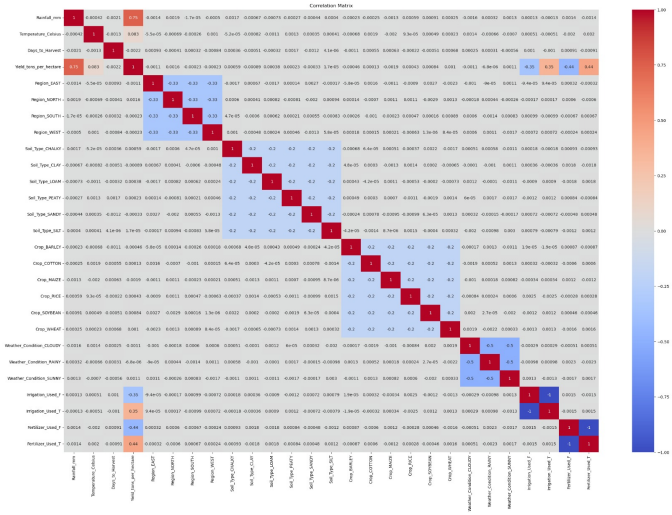
Bar plot titled "Average Yield by Region" shows very similar average yields across all four regions (East, North, South, and West), with averages around 4.5 tons per hectare.

I. Relationship between fertilizer usage and yield, separated by irrigation



The faceted scatter plots explore the relationship between Fertilizer_Used, Irrigation_Used, and Yield_tons_per_hectare. It appears that using fertilizer (Fertilizer_Used = T) generally leads to higher yields, regardless of irrigation.

J. Correlation matrix of continuous variable



The correlation matrix reveals a strong positive correlation between Days_to_Harvest and Yield_tons_per_hectare (0.75), confirming the earlier observation from the pair plot. Additionally, some moderate negative correlations are observed between specific one-hot encoded columns, likely due to the nature of categorical variables.

VI. Machine Learning Algorithms

A. Linear Regression

- **Justification:** Linear regression is always considered 1st choice when it comes to applying regression model due to its simplicity. This regression model assumes linear relation between dependent variables and independent variables. In this project the target variable 'crop_yield_per_hecture' may depend linearly on some of the feature variables 'Rainfall_mm', 'Temperature_Celsius', 'Soil_Type'. Thus, this regression model was considered.
- **Model Training and Tuning:** The data was scaled to standardize the features, ensuring that the coefficients learned by the model were balanced. No hyperparameters were used as Linear regression does not require any.
- **Effectiveness:** The model performed well with an mean squared error value is 0.336, the accuracy of the model is 88.27%. Therefore, this model can be used to efficiently predict the target variable.

B. K-NN Regression

- **Justification:** This model is used to capture the non-linear relationship between independent variables and dependent variable. Since, we can't rely on the assumption that there will only be linear relationship between the features and target we have to also rely on the possibility of non-linear relationship. KNN is a simple, interpretable since predictions are based on the nearest neighbors and lazy-learning algorithm.
- **Model Training and Tuning:** The value of K is set to 5 while implementing the model. Future tuning could involve changing the value of K if the accuracy is not satisfactory.
- **Effectiveness:** The model performed well with an mean squared error value of 0.432 and the accuracy of the model is 84.38%. Therefore, this model can also be used to efficiently predict the target variable.

C. Naïve Bayes Classification

- **Justification:** This algorithm was implied because it uses conditional probability to classify data. Implementing this model can provide useful information regarding target variable, if we divide the target variable into two categories i.e Low Yield, High Yield. The model can classify in these two categories provided the values of the features. It is a simple, effective algorithm especially when dealing with categorical data.
- **Model Training and Tuning:** The model was straightforward to train, and no parameter tuning was necessary.
- **Effectiveness:** The model performed well with an precision score of 0.8938, f1_score of 0.8938, recall score of 0.8938. Therefore, this model can also be used to efficiently predict the target variable.

D. Decision Tree Regression:

- **Justification:** This model was implemented due ti its simplicity in finding the non-linear relationship between features and target and its ability to model interaction between different features. Also, in this model feature scaling is not required therefore, this can simplify the preprocessing stage. This model provides faster predictions as compared to other models.
- **Model Training and Tuning:** The decision tree was trained with a default depth, and no extensive tuning was performed.
- **Effectiveness:** The model performed well with an mean square error value of 0.710, R2 value of 0.752. Therefore, this model can also be considered to efficiently predict the target variable.

E. Random Forest Regression:

- **Justification:** This model is implemented because it captures the non-linearities between features, target and it can reduce overfitting and provide more accurate results by aggregating multiple decision trees.
- **Model Training and Tuning:** The model was trained with 100 trees, and no extensive hyperparameter tuning was performed initially.
- **Effectiveness:** The model performed well with a mean square error value of 0.358, R2 value of 0.875. Therefore, this model can also be considered to efficiently predict the target variable.

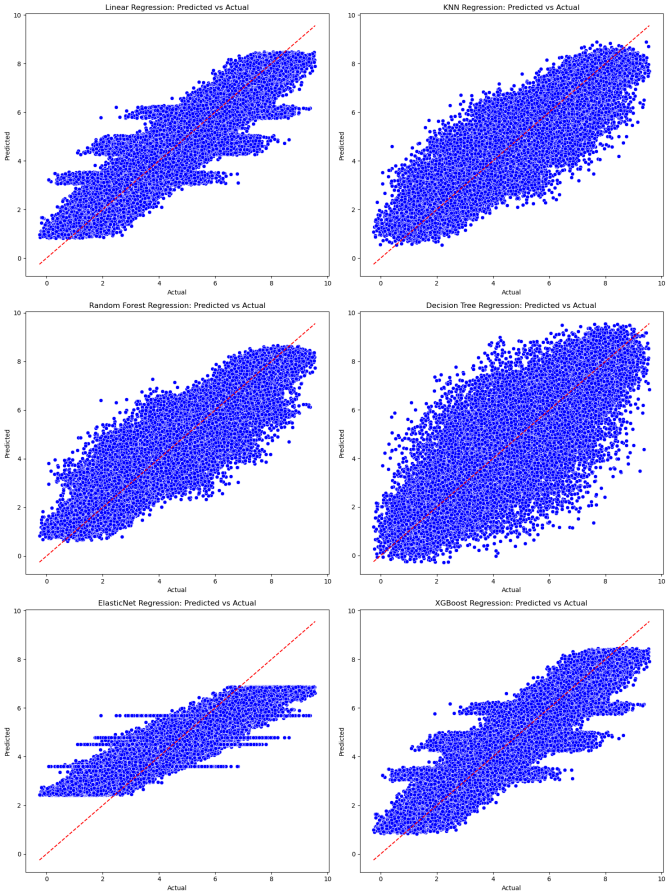
F. Elastic Net Regression:

- **Justification:** This model combines L1(LASSO) and L2(Ridge) regression models so that the data becomes suitable for many features. It prevents overfitting while still allowing feature selection.
- **Model Training and Tuning:** The alpha parameter is set to 0.1, l_1 ratio to 0.5. These parameters can be further tuned for better regularization.
- **Effectiveness:** The model performed well, with a mean square error value of 0.7618, R2 value of 0.738. It provides a balance between bias and variance, with the regularization term helping to reduce overfitting.

G. XG Boost Regression:

- **Justification:** This model was implemented because it can efficiently handle large and complex datasets and non-linear relationships.
- **Model Training and Tuning:** The learning rate and maximum depth were set to default values of 0.1 and 5, respectively.
- **Effectiveness:** The model performed well with a mean square error value of 0.337 and R2 value of 0.882. Therefore, this model can also be considered for efficiently predicting the target variable.

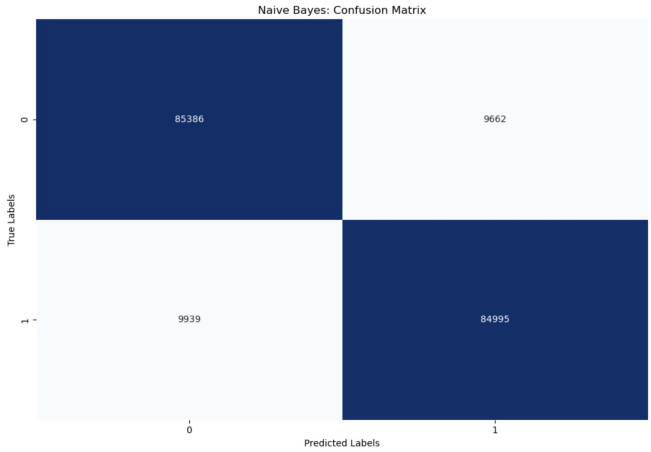
H. Machine Learning Models Visualization



Actual data value V/S Predicted data value Plot

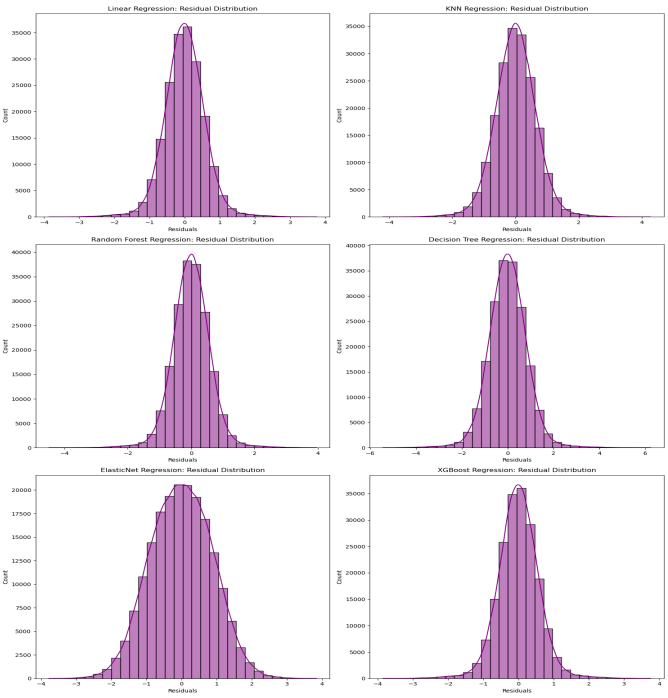
- **Linear Regression:** The data points are spread around the diagonal line, but some data points are scattered. This implies that model does not perform well on complex relationships in the data, which leads to underfitting.
- **KNN Regression:** Graph indicates more data points are spread around the diagonal line as compared to Linear Regression which implies that KNN handled complex relationships better than Linear Regression model but, still some data points are scattred.
- **Random Forest Regression:** The number of data points clustered around the diagonal line is more as compared to both Linear and KNN regression. Further, the scattering of data points is also tighter as compared to Linear and KNN regression models. This implies that Random Forest Regression handles complex relationships better.
- **Decision Tree Regression** The data points are scattered more as compared to Random Forest Regression model. This might overfit the data, leading to poorer generalization.
- **Elastic Net Regression:** There are some visible levels in the graph that clearly shows the regularization imposed. This might lead to less flexibility. This led to improper capturing of certain relationships.
- **XGBoost Rgression:** The data points are clustered around the diagonal line with less scattering but not as

less as Random Forest Regression. This model handles complex relationships quite well too.



Naive Bayes Classification Confusion Matrix Plot

- **Naive Bayes Classification:** 84,995 data points belong to True Positive which indicate instances where actual label was 1 and model predicted 1. 85,386 data points belong to True Negative which indicate instances where actual label was 0 and model predicted 0. 9,662 data points belong to False Positive which indicate instances where actual label was 0 and model predicted 1. 9939 data points belong to False Negative which indicate instances where actual label was 1 and model predicted 0. Therefore, this shows that model correctly predicted many data points as True Positive, False Negative respectively indicating accurate performance of the model.



Residual Plot of Regression models

- **Analyzation:** The residual plots of all the graphs are symmetrical around 0 which indicates good efficiency of the models. Though, some slight spread is observed across all the graphs, the minimum spread is of Random Forest model. Indicating that the model handles complex relationships better as compared to others.

I. Most Suited Machine Learning Model:

According to the problem statement, we have to predict crop yeild production in tons per hectare based on environmental and agricultural conditions. Therefore, we have to choose a regression model. Thus, the most suited regression model according to the accuracy, visualizations would be Random Forest Regression. Here are some advantages of this model over others :-

- **Handling Complex Relationships:** The problem contains several features including temperature, rainfall, soil type. Random Forest Regression model excels in capturing complex relationships among them making it ideal as compared to others.
- **Robust Performance:** Upon implementation of this model, it produced very efficient mean squared error value of 0.358, R^2 value of 0.875. The visualizations show that this model has most number of data points clustered around the diagonal line and least spread of residuals.
- **Generalization:** Random Forest models are generally robust against overfitting since, they average predictions from multiple decision trees which can also be visualized from Residual plot, Predicted V/S Actual data points plot graphs. This suggests that the model generalizes very well to the new data provided. In real world applications, this factor plays a very vital importance especially in agriculture where conditions, temperature, soil type, rainfall and many more factors differ according to region, weather season.

Distributed Data Cleaning/Processing

REFERENCES

- https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield
  Dataset is clean, had to make it unclean using python script, before performing data cleaning operations.
- https://ieeexplore.ieee.org/document/9432236/figuresfigures
- https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222
- Links used to refer code:
  1. https://numpy.org/doc/
  2. https://pandas.pydata.org/docs/
  3. https://seaborn.pydata.org
  4. https://matplotlib.org/stable/index.html
  5. https://www.geeksforgeeks.org
  6. https://www.geeksforgeeks.org/python-linear-regression-using-sklearn/
  7. https://www.geeksforgeeks.org/k-nearest-neighbors-with-python-ml/

8. https://www.geeksforgeeks.org/ml-naive-bayes-scratch-implementation-using-python/
9. https://www.geeksforgeeks.org/random-forest-regression-in-python/
10. https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/
11. https://www.geeksforgeeks.org/what-is-elasticnet-in-sklearn/
12. https://www.geeksforgeeks.org/xgboost-for-regression/