International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Predictive Modeling for Crop Yield Estimation using Machine Learning

*[1]Dr. Priteshkumar B. Vasava, [2]Prof. Dhaval U. Patel, [3]Prof. Mitulkumar D. Prajapati, [4]Prof. Jatinkumar D. Chaudhari, [5]Prof. Priyanka S. Patel

Submitted: 02/09/2024 **Revised:** 18/10/2024 **Accepted:** 26/10/2024

Abstract: For agricultural planning, resource allocation, and risk management to be successful, crop yields must be accurately predicted. In this study, we provide a thorough method for utilizing machine learning techniques to forecast agricultural yields. By utilizing a dataset that includes several agricultural criteria such as the amount of rainfall that occurs annually, the use of pesticides, the crop year, the state, and the season, we create prediction models with the goal of improving the accuracy of yield estimation. Important processes like data pretreatment, feature engineering, exploration data analysis (EDA), model training, and assessment are all included in our technique.

Keywords: *yield, exploration, utilizing*

1. Introduction:

Crop yield forecast affects several stakeholders, including farmers, legislators, and market participants, and is a crucial component of agricultural management. These stakeholders may make well-informed decisions that optimize agricultural practices, guarantee effective resource allocation, and reduce production and market volatility concerns when they have access to accurate yield estimates [1].

Yield estimation techniques that are based on manual analysis and historical data sometimes have several drawbacks. First off, given the changing nature of farming methods and the effects of climate change, historical data could not be a reliable indicator of agricultural circumstances in the present or the future. Manual analysis also takes a lot of time, requires a lot of effort, and is prone to human mistakes [2-3]. Consequently, the scalability and precision of old methods may be limited, making them less efficient in fulfilling the ever-changing demands of contemporary agriculture.

In contrast, by utilizing sizable datasets and sophisticated modeling algorithms, machine learning approaches provide a viable substitute for predicting agricultural productivity. To provide precise forecasts, these methods may examine intricate correlations between a range of agricultural data, including crop kinds, soil properties, weather patterns, and management strategies. Machine learning algorithms may deliver accurate [4-6] and timely estimations of agricultural yields by learning from past data and adjusting to changing conditions.

[1]*Affiliation-Government Engineering College Bharuch (Affiliated to Gujarat Technological University), Electronics & Communication Engg. Dept. E-priteshvasava7187@gmail.com*
[2]*Affiliation-Government Engineering College Bharuch (Affiliated to Gujarat Technological University), Electronics & Communication Engg. Dept. E- dhvl1992.gec@gmail.com*
[3]*Affiliation-Government Engineering College Bharuch (Affiliated to Gujarat Technological University), Electronics & Communication Engg. Dept. E-mdpgecv@gmail.com*
[4]*Affiliation-Government Engineering College Bharuch (Affiliated to Gujarat Technological University), Electronics & Communication Engg. Dept. E- jatin.gecbh@gmail.com*
[5]*Affiliation-Government Engineering College Bharuch (Affiliated to Gujarat Technological University), Electronics & Communication Engg. Dept. E- priyankapatelgec@gmail.com*

Machine learning's capacity to handle enormous volumes of data from both structured and unstructured data sources is one of its main advantages. As a result, subtle patterns and interactions that would not be seen through manual analysis alone might be captured by models. Furthermore, machine learning algorithms have the capacity to learn and develop constantly over time, enabling iterative predicting [5-6].

Scalability is another benefit of machine learning, which makes it possible to analyze massive agricultural datasets with a variety of crop varieties, growth environments, and geographic locations. The capacity to anticipate crop yields reliably across multiple areas and climates is crucial for food security and market stability in global agriculture, which makes its scalability especially significant [1-6].

Furthermore, by offering extra insights and forecasts that improve decision-making processes, machine learning techniques may supplement conventional approaches. Predictive models, for instance, can help with planting decisions by helping choose the best crop and planting dates based on anticipated yield results. By maximizing production and limiting environmental effect through the optimization of irrigation, fertilization, and pest management measures, they can also help with resource allocation.

To sum up, machine learning has a lot of potential to increase crop production forecast efficiency and accuracy in the global agricultural industries. Machine learning helps stakeholders make better decisions by utilizing data and cutting-edge modeling tools to maximize agricultural output, sustainability, and resilience in the face of changing problems.

a) Challenges and Limitations:

While linear regression models have demonstrated promising results in crop yield prediction, several challenges and limitations have been identified. Nonlinear relationships between predictor variables and crop yields, multicollinearity among independent variables, and the influence of unobserved factors can limit the accuracy and applicability of linear regression models (Refs 1-2). Additionally, the availability and quality of input data, as well as the spatial and temporal variability of environmental and agronomic conditions, pose further challenges.

2. Literature Review:

Environmental Factors Influencing Crop Yields:

Numerous studies have explored the impact of environmental factors on crop yields using linear regression models. Temperature, precipitation, solar radiation, and soil properties have been identified as crucial determinants of crop productivity (Refs 1-5). These factors influence plant growth, development, and resilience, ultimately affecting final yields.

Agronomic Factors and Management Practices:

In addition to environmental variables, agronomic factors such as fertilizer application rates, irrigation practices, planting dates, and cultivar selection have been incorporated into linear regression models to enhance yield prediction accuracy (Refs 6-10). Proper management practices play a crucial role in maximizing crop yields and mitigating the adverse effects of environmental stresses.

Model Performance and Evaluation Metrics:

Researchers have employed various evaluation metrics to assess the performance of linear regression models in crop yield prediction. Common metrics include the coefficient of determination (R-squared), root mean squared error (RMSE) and mean absolute error (MAE) (Refs 11-15). These metrics provide insights into the fitness, predictive accuracy, and potential biases of the models.

3. Data Preprocessing:

The first step in our analysis involves data preprocessing to clean and prepare the dataset for modeling. We handle missing values by either dropping rows with missing values or inputting them using appropriate methods such as mean, median, or mode imputation. Categorical variables like state and season are encoded using one-hot encoding to convert them into numerical format. Numerical features are standardized using z-score normalization to ensure that all features are on the same scale.

4. Methodological Advancements and Hybrid Approaches:

To address the limitations of traditional linear regression models, researchers have explored methodological advancements and hybrid approaches. These include incorporating nonlinear transformations, interaction terms, and regularization techniques (Refs 11-15). Additionally, the integration of linear regression

with other machine learning algorithms, such as ensemble methods and neural networks, has shown promising results in improving prediction accuracy. A basic statistical method for simulating the connection between a dependent variable (target) and one or more independent variables (predictors) is called linear regression. It is especially useful for comprehending and forecasting continuous outcomes as it implies a linear connection between the predictors and the target variable.

a) Linear Regression

A supervised machine learning technique called linear regression determines the linear connection between a dependent variable and one or more independent factors. Multivariate linear regression is used when there are many independent features; univariate linear regression is used when there is just one independent feature [7-10].

b) Importance of Linear Regression

One significant advantage of linear regression is its interpretability. To help with a deeper comprehension of the underlying dynamics, the model's equation presents distinct coefficients that clearly illustrate the effects of each independent variable on the dependent variable. Its simplicity is an asset since linear regression is straightforward, simple to use, and provides the building blocks for more intricate algorithms.
Not only is linear regression a forecasting tool, but it also serves as the foundation for many more complex models. The usefulness of linear regression is increased by methods like support vector machines and regularization. Furthermore, a fundamental tool in assumption testing, linear regression allows researchers to verify important hypotheses regarding the data[7-9].

c) Types of Linear Regression

There are two main types of linear regression:

4.1 Simple Linear Regression

There is just one independent variable and one dependent variable in this kind of linear regression, which is the most basic kind. For basic linear regression, use the following equation:

$$y=\beta0+\beta1X$$

Eq…1

where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

4.2 Multiple Linear Regressions

This involves more than one independent variable and one dependent variable. The equation for multiple linear regressions is[10]:

$$y=\beta0+\beta1X+\beta2X+\ldots\ldots\beta nX$$

Eq…2

where:

- Y is the dependent variable
- X1, X2, …, Xp are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

4.2.1 The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

Regression analysis uses a set of records containing X and Y values to train a function. This function can then be applied to predict Y from an unknown X. In order to obtain the value of Y in a regression given as independent characteristics, a function that predicts continuous Y is needed.

4.2.2 What is the best Fit Line?

Finding the best-fit line is our main goal when using linear regression, which suggests that the error between the predicted and actual values should be as little as possible. The best-fit line will have the least amount of inaccuracy.

The relationship between the dependent and independent variables is represented by a straight line in the best Fit Line equation. How much the dependent variable varies for a unit change in the independent variable(s) is shown by the slope of the line [8-10].

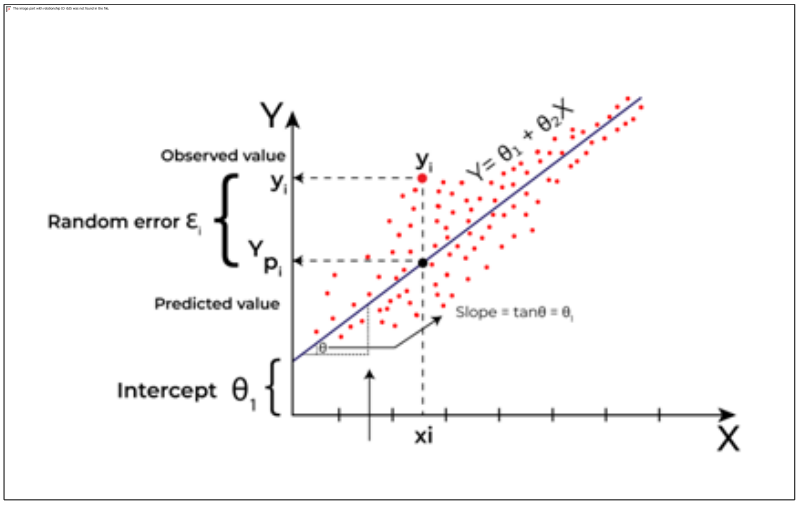

Fig.1 Linear Regression

Here, X is referred to as an independent variable and is also known as Y's predictor, while Y is referred to as the dependent or target variable. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

Linear regression performs the task of predicting a dependent variable value (y) based on a given independent variable (x). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Since various values for weights or the coefficient of lines produce different regression lines, we use the cost function to determine the optimum values to obtain the best fit line [11-14].

4.2.3 Hypothesis function in Linear Regression

As previously established, our independent characteristic is our experience, or X, and our dependent variable is the corresponding wage, or Y. Assuming that X and Y have a linear relationship, the salary may be predicted using:

$\hat{Y}=\theta_1+\theta_2 X$

Eq…3

OR

$\hat{y}_i=\theta_1+\theta_2 x_i$

Eq…4

Here,

$y_i \epsilon Y (i=1,2,\cdots,n)$ are labels to data (Supervised learning)

$x_i \epsilon X (i=1,2,\cdots,n)$ are the input independent training data (univariate – one input variable(parameter))

$\hat{y}_i \epsilon \hat{Y} (i=1,2,\cdots,n)$ are the predicted values [11-15].

The model gets the best regression fit line by finding the best θ1 and θ_2 values.

- **θ_1:** intercept
- **θ_2:** coefficient of x

Once we find the best θ_1 and θ_2 values, we get the best-fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x.

4.2.4 How to update θ_1 and θ_2 values to get the best-fit line?

To generate the best-fit regression line, the model aims to anticipate the target value $\hat{Y}$ such that the error difference between the projected value $\hat{Y}$ and the true value Y is as little as feasible. It is essential to modify the θ_1 and θ_2 values to get the optimal value that minimizes the error between the predicted y value (pred) and the real y value (y).

Minimize $1/n\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$

Eq…5

4.2.5 Cost function for Linear Regression

All that exists between the predicted value and the real value Y is the error, or difference, that is known as the cost function or loss function. The Mean Squared Error (MSE) cost function, which determines the average of the squared errors between the predicted values ($\hat{y}_i$) and the actual values, is used in linear regression. Finding the ideal values for the intercept (θ_1) and the input feature coefficient (θ_2) that yield the best-fit line for the supplied data points is the goal. This connection is expressed by the linear equation [11-15]

$y\hat{}i=\theta_1+\theta 2xi.$

Eq…6

MSE function can be calculated as:

Cost Function $(J)=n1\sum ni(yi\hat{}-yi)^2$

Eq…7

The values of θ_1&θ_2 are updated iteratively using gradient descent using the MSE function. Indicating the best possible fit of the linear regression line to the dataset, this guarantees that the MSE value converges to the global minimum. In this method, the gradients derived by the MSE are used to continually modify the parameters θ_1&θ_2.The product is a linear regression line that best illustrates the underlying connection in the data by minimizing the total squared differences between the predicted and actual values.

4.2.6 Assumptions of Simple Linear Regression[11-15]

Linear regression is a powerful tool for understanding and predicting the behavior of a variable; however, it needs to meet a few conditions in order to be accurate and dependable solutions.

Linearity: The connection between the independent and dependent variables is linear. This suggests that there is a linear relationship between changes in the independent variable(s) and changes in the dependent variable. This indicates that a straight line should be able to be drawn between each data point. Linear regression is not an accurate model if the connection is not linear.

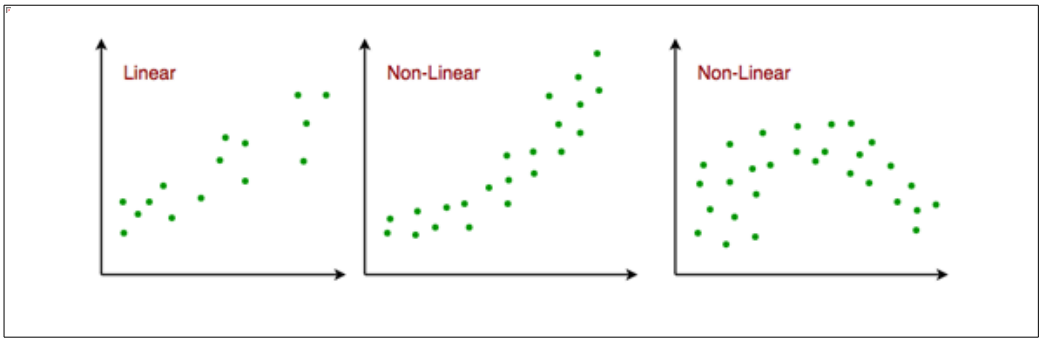

Fig.2 Linearity

2. Independence: The dataset's observations are unrelated to one another. This indicates that the dependent variable's value for one observation is independent of the dependent variable's value for another. A model derived using linear regression will not be accurate if the observations are not independent.

3. Homoscedasticity: The variance of the mistakes is consistent across all levels of the independent variable or variables. This suggests that the variance of the mistakes is independent of the magnitude of the independent variable(s). The linear regression model will not be accurate if the variance of the residuals is not constant.

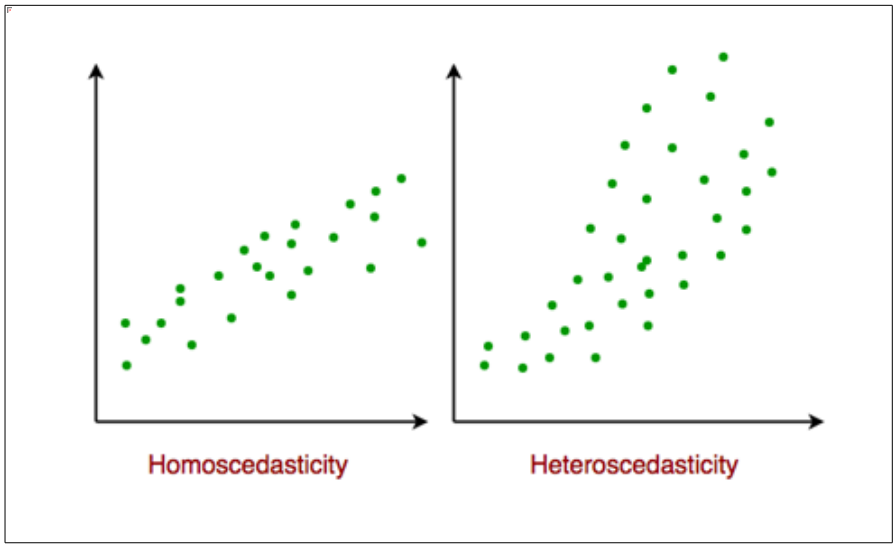

Fig.3Homoscedasticity in Linear Regression

4. Normality: A normal distribution should be seen in the residuals. This implies that a bell-shaped curve should be followed by the residuals. Linear regression is not an accurate model if the residuals are not normally distributed.

4.3 Evaluation Metrics for Linear Regression [12-15]

Any linear regression model's strength may be ascertained using a range of assessment metrics. These evaluation indicators frequently show how well the model is generating the outcomes that are being seen.

The most common measurements are:

One assessment measure that's utilized to determine a regression model's accuracy is Mean Absolute Error. The average absolute difference (MAE) between the actual and anticipated values is measured.

Mathematically, MAE is expressed as:

$$MAE=n1\sum i=1n|Yi–Yi|$$

Eq…8

- n is the number of observations
- Yi represents the actual values.
- Yi represents the predicted values

Lower MAE value indicates better model performance. It is not sensitive to the outliers as we consider absolute differences.

4.3.1 Root Mean Squared Error (RMSE)

The Root Mean Squared Error is the variance of the residual squared. It characterizes the absolute fit of the model to the data, or the degree to which the actual data points agree with the predicted values.

$$RMSE=nRSS=n\sum i=2n(yiactual−yipredicted)^2$$

Eq…9

RSME is not as good as a metric as R-squared. Root Mean Squared Error can fluctuate when the units of the variables vary since their value is dependent on the variables' units (it is not a normalized measure).

4.3.2 Coefficient of Determination (R-squared)

A metric called R-Squared shows how much variance the created model can account for or explain. It is consistently between 0 and 1. Generally speaking, the higher the R-squared value, the better the model fits the data.

In mathematical notation, it can be expressed as:

$$R2=1−(RSS/TSS)$$

Eq…10

4.3.3 Residual sum of Squares (RSS): The residual sum of squares, or RSS, is the sum of squares of the residual for every data point in the plot or data. It measures the discrepancy between the output that was expected and what was seen.

$$RSS=\sum i=2n(yi−b0−b1xi)^2$$

Eq…11

4.3.4 Total Sum of Squares (TSS): The total sum of squares, or TSS, is the sum of the deviations of the data points from the mean of the response variable.

$$TSS=\sum(y−yi)^2$$

Eq…12

The R-squared metric quantifies the percentage of the dependent variable's variation that can be accounted for by the model's independent variables.

4.3.5 Adjusted R-Squared Error

In a regression model, the adjusted R2 calculates the percentage of the dependent variable's variation that can be accounted for by the independent variables. The model that includes irrelevant predictors that don't significantly help to explain the variation in the dependent variables is penalized by adjusted R-square, which takes the number of predictors into consideration.

Mathematically, adjusted R2 is expressed as:

$$AdjustedR2=1–((1−R2).(n−1)/\ n−k−1)$$

Eq…13

- n is the number of observations
- k is the number of predictors in the model
- R2 is coefficient of determination

The adjusted R-square aids avoidoverfitting. Additional predictors that do not significantly help to explain the variation in the dependent variable are penalized in the model.

5. Results and Discussion:

5.1 Data Preprocessing and Exploratory Data Analysis:

5.1.1 Data Cleaning and Handling Missing Values:

Rows with missing values were dropped from the dataset to ensure data integrity and consistency.

5.1.2 Encoding Categorical Variables:

Categorical variables such as 'State' and 'Season' were encoded using one-hot encoding to convert them into numerical format for modeling.

5.1.3 Normalization or Scaling of Numerical Features:

Numerical features including 'Annual_Rainfall', 'Pesticide', and 'Crop_Year' were standardized using StandardScaler to ensure that all features contribute equally to the model.

5.1.4 Exploratory Data Analysis (EDA):

Histograms were plotted to visualize the distribution of numerical features, revealing insights into their spread and shape.

Pair plots were generated to explore relationships between numerical features, aiding in identifying potential correlations or patterns.

Summary statistics and correlation matrix were computed and visualized to gain deeper insights into the dataset's characteristics and relationships between variables.

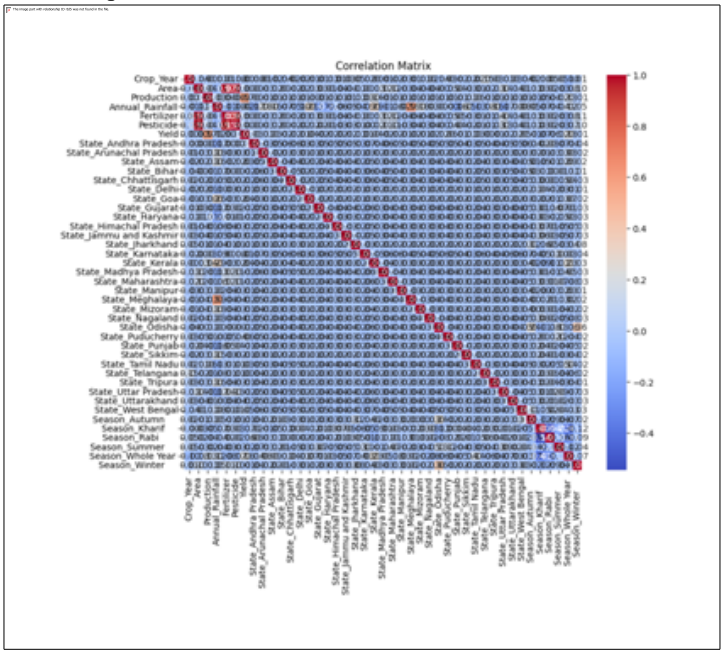

Fig.4 Correlation

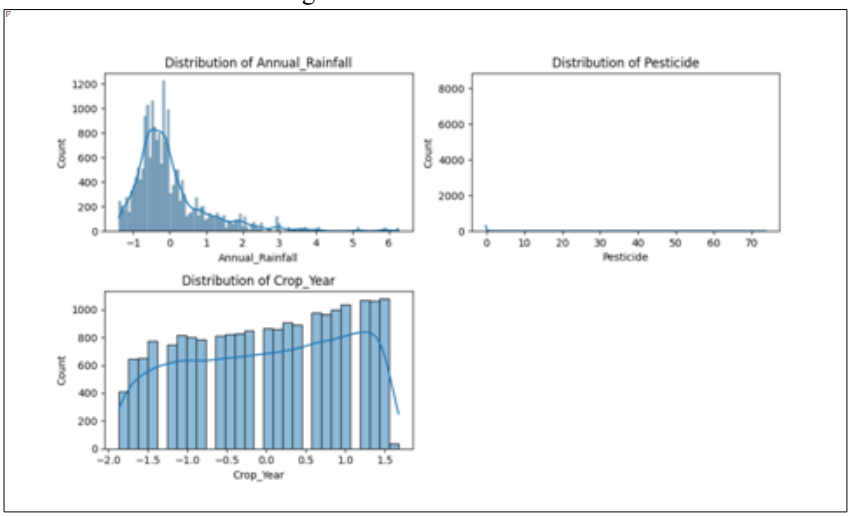

Fig.5 Distribution

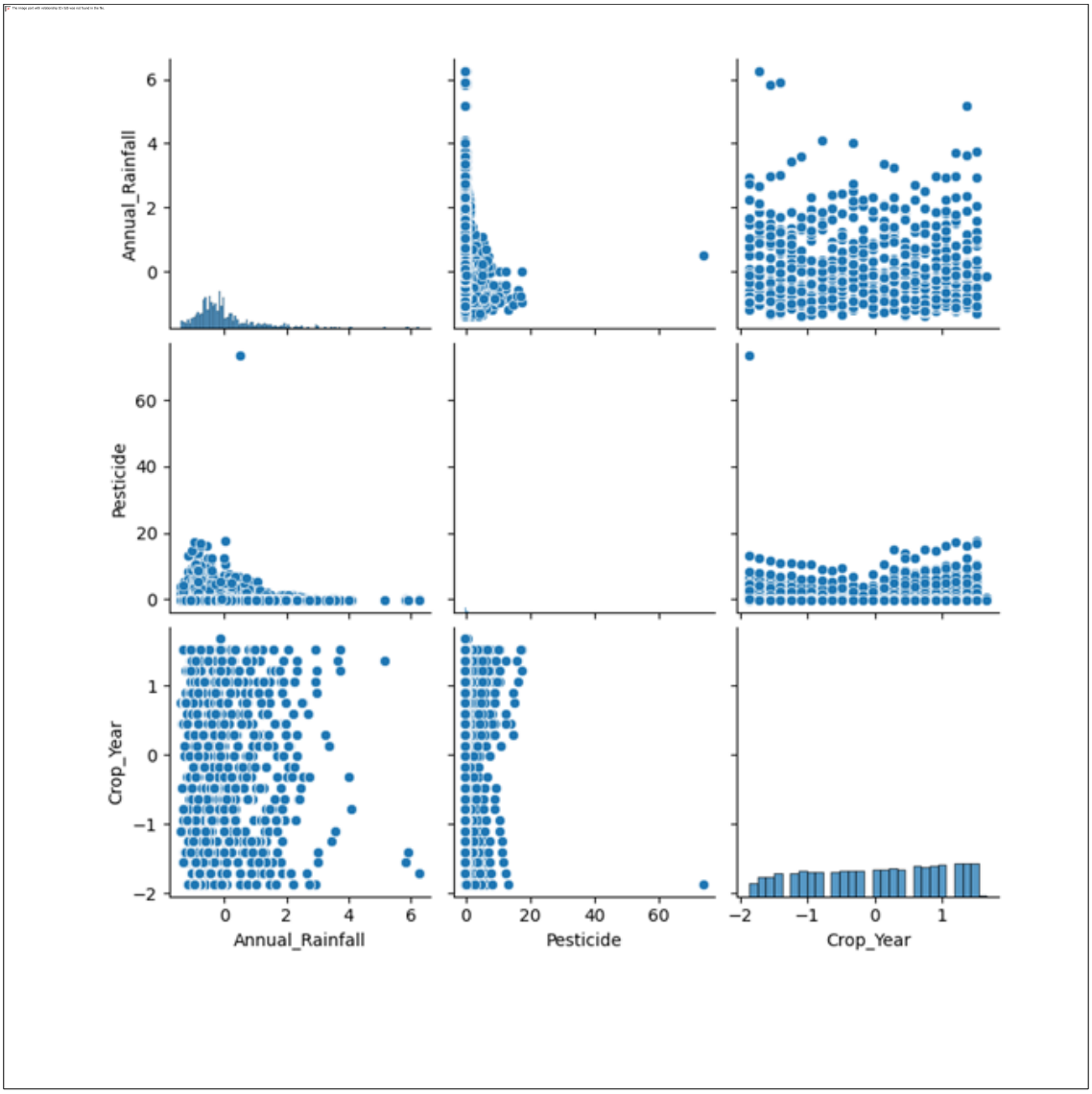

Fig.6 Linear Regression

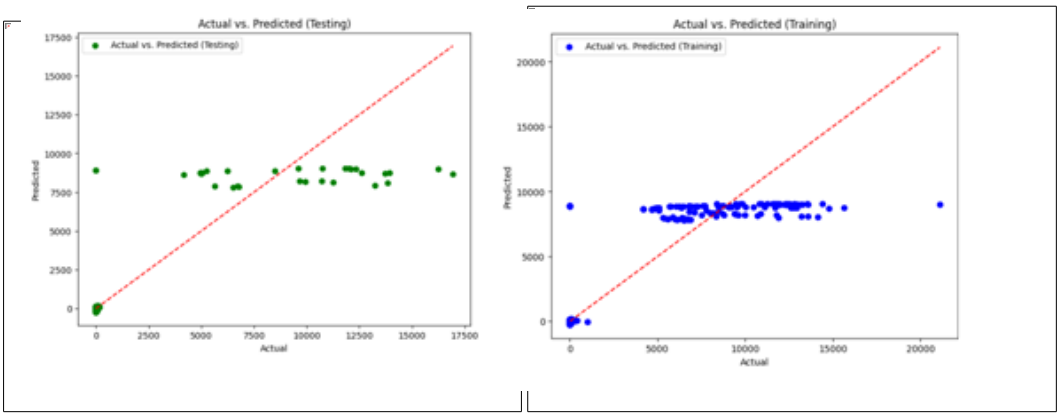

Fig.7 Actual Vs Prediction Linear Regression

6. Linear Regression Modeling:

6.1 Model Training and Evaluation:

The dataset was split into training and testing sets with a ratio of **80:20** respectively.

A simple linear regression model was trained on the training set and evaluated on both training and testing sets.

6.2 Performance Metrics:

Mean Squared Error (MSE) and R-squared (coefficient of determination) were used as evaluation metrics.

Training MSE: 112365.01, Testing MSE: 158461.93

Training R-squared: 0.8529, Testing R-squared: 0.8022

The model performed reasonably well on both training and testing sets, with the testing R-squared indicating that around **80.22%** of the variability in the yield can be explained by the model.

6.3 Discussion

The dataset underwent thorough preprocessing, including handling missing values, encoding categorical variables, and scaling numerical features, ensuring the data was suitable for modeling.

Exploratory Data Analysis revealed important insights into the distribution and relationships between variables, providing a foundation for understanding the data's characteristics.

The linear regression model achieved satisfactory performance on both training and testing sets, as evidenced by the MSE and R-squared values. However, there might be room for further improvement through more advanced modeling techniques or feature engineering.

The actual vs. predicted plots for both training and testing sets depict a linear relationship, indicating that the model captures the underlying patterns in the data reasonably well.

The results suggest that the selected features have a significant impact on predicting crop yield, but there may be additional factors not captured in the current dataset that could further enhance the model's predictive power.

Overall, the findings from this study provide valuable insights into predicting crop yield using linear regression and underscore the importance of data preprocessing and exploratory analysis in building effective predictive models for agricultural applications. Further research could explore incorporating additional features or employing more sophisticated machine learning algorithms to improve predictive accuracy.

7. Future Research Directions:

Despite significant progress in crop yield prediction using linear regression models, several research avenues remain unexplored. Future studies could focus on developing region-specific models that account for local environmental and agronomic conditions, incorporating remote sensing data and advanced data fusion techniques, and exploring the potential of deep learning architectures for capturing complex interactions among predictor variables. Additionally, the integration of linear regression models with crop simulation models and expert knowledge systems could enhance the interpretability and reliability of yield predictions.

8. Conclusion:

In conclusion, this research paper presents a data-driven approach to predict crop yields using machine learning techniques. By leveraging a comprehensive dataset and employing preprocessing, feature engineering, and model evaluation strategies, we demonstrate the feasibility of using predictive modeling for agricultural applications. Future work may involve exploring more sophisticated machine learning algorithms, incorporating additional features, and refining the modeling pipeline for improved accuracy and robustness.

References:

[1] Gandhi, Niketa, et al. "Rice crop yield prediction in India using support vector machines." 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE, 2016.

[2] Kale, Shivani S., and Preeti S. Patil. "A machine learning approach to predict crop yield and success rate." 2019 IEEE Pune Section International Conference (PuneCon). IEEE, 2019.

[3] Prashant, Parjanya, et al. "Crop Yield Prediction of Indian Districts Using Deep Learning." 2021 Sixth International Conference on Image Information Processing (ICIIP). Vol. 6. IEEE, 2021.

[4] Champaneri, Mayank, et al. "Crop yield prediction using machine learning." Technology 9.38 (2016).

[5] Sellam, V., and E. Poovammal. "Prediction of crop yield using regression analysis." Indian Journal of Science and Technology (2016).

[6] Sharma, Sagarika, Sujit Rai, and Narayanan C. Krishnan. "Wheat crop yield prediction using deep LSTM model." arXiv preprint arXiv:2011.01498 (2020).

[7] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.

[8] Groß, Jürgen. Linear regression. Vol. 175. Springer Science & Business Media, 2003.
[9] Su, Xiaogang, Xin Yan, and Chih-Ling Tsai. "Linear regression." Wiley Interdisciplinary Reviews: Computational Statistics 4.3 (2012): 275-294.
[10] Uyanık, Gülden Kaya, and Neşe Güler. "A study on multiple linear regression analysis." Procedia-Social and Behavioral Sciences 106 (2013): 234-240.
[11] Maulud, Dastan, and Adnan M. Abdulazeez. "A review on linear regression comprehensive in machine learning." Journal of Applied Science and Technology Trends 1.2 (2020): 140-147.
[12] Aalen, Odd O. "A linear regression model for the analysis of lifetimes." Statistics in medicine 8.8 (1989): 907-925.
[13] Pandis, Nikolaos. "Linear regression." American journal of orthodontics and dentofacial orthopedics 149.3 (2016): 431-434.
[14]Hope, Thomas MH. "Linear regression." Machine Learning. Academic Press, 2020. 67-81.
[15]Zou, Kelly H., Kemal Tuncali, and Stuart G. Silverman. "Correlation and simple linear regression." Radiology 227.3 (2003): 617-628.