Cultivating Forecasts: Unveiling Crop Yield Prediction with Linear Regression

# Arun Prasad S, Dr.Aghila R

***Abstract*:** Machine learning has emerged as a pivotal research area in agriculture, particularly in the analysis and prediction of crop yields. With the ever-increasing complexity of agricultural data, deciphering valuable underlying patterns becomes challenging for traditional approaches. However, by harnessing machine learning strategies, we can automatically access and comprehend these intricate patterns. In this study, we investigate the predictability of pesticide usage in various countries using the quantity of pesticides used annually from 1990 to 2016 through the implementation of a linear regression model. This research focuses on predicting pesticide usage for crop yield using a comprehensive agricultural dataset collected from kaggle.com. Our methodology involves employing regression analysis to assess the accuracy and effectiveness of yield predictions for crops with the quantity of pesticides used in various countries. We establish relationships between variables such as value

tonnes of active ingredients used and year of crop yield. The implications of this study are significant for farmers who seek to conceptualize their expected yields during the growing season. By accurately measuring the potential pesticide usage from different years, farmers can make informed decisions and mitigate losses. The financial impact of crop yield on farmers is considerable, making the predictive power of our regression model a valuable tool for avoiding potential setbacks. Throughout this research paper, we meticulously examine the accuracy of our regression-based predictions, which has broader implications for enhancing crop yield forecasting in the agricultural sector.

***Keywords: Machine learning,******Linear regression, R Square****, RMSE*

## I.INTRODUCTION

The agriculture sector plays a critical role in sustaining global food security and economic growth. In recent years, the advent of machine learning has revolutionized agricultural research, offering promising solutions for analyzing and predicting crop yields. By harnessing the power of machine learning strategies, researchers have been able to delve into complex agricultural data and uncover valuable underlying patterns that were previously challenging to decipher using traditional approaches.

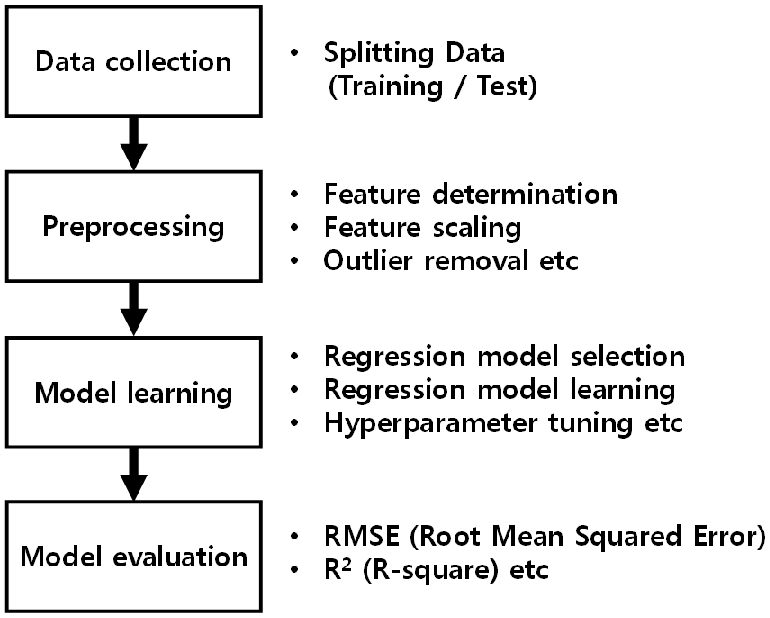
One of the essential factors affecting crop productivity is the usage of pesticides. Pesticides are crucial in safeguarding crops from pests and diseases, ensuring optimal yield and quality. However, determining the appropriate quantity of pesticides to be used in different countries and under varying

physical conditions remains a challenging task.

In this project, we aim to explore the predictability of pesticide usage in various countries by leveraging the insights provided by machine learning techniques. Our focus is on analyzing the quantity of pesticides used annually from 1990 to 2016, as recorded in a comprehensive agricultural dataset obtained from kaggle.com. To achieve this, we employ a linear regression model, which allows us to establish relationships between the quantity of pesticides used and the corresponding crop yield for different years and countries.

By predicting pesticide usage for crop yield, our research intends to assist farmers in making informed decisions during the growing season. Accurate predictions of pesticide requirements can lead to optimized agricultural practices, ultimately reducing losses and maximizing productivity. Considering the significant financial impact that crop yield has on farmers' livelihoods, the outcomes of this study hold potential for preventing potential setbacks and enhancing overall agricultural sustainability.

Throughout this project, we meticulously examine the accuracy and effectiveness of our regression-based predictions. Moreover, our findings have broader implications for enhancing crop yield forecasting in the agricultural sector, showcasing the potential of machine learning in transforming the way we approach and address complex agricultural challenges.



### Fig. 1.Machine leaning process

If you want to do prediction or forecasting using machine learning algorithms, then you must follow the basic steps presented in fig.1. In the data collection step, there is the various source by which we can get data. Next, the data is bifurcated by preprocessing methods, in which categorical and null values are handled. Appropriate learning algorithms is to be selected according to build the model. And the last, results are evaluated using RMSE,R Square and chart is prepared using data visualization tools

## II.OBJECTIVES

The primary objective of this research project is to explore the feasibility of using machine learning techniques, specifically linear regression, to predict pesticide usage in various countries for crop yield analysis. To achieve this overarching goal, the project aims to accomplish the following specific objectives:

1. Analyze Pesticide Usage Data: Collect and preprocess a comprehensive dataset containing historical records of pesticide usage in different countries from 1990 to 2016. This dataset will serve as the foundation for training and evaluating the regression model.

2. Implement Linear Regression Model: Develop and deploy a linear regression model to establish relationships between the quantity of pesticides used and crop yield data from the available dataset. The model will identify patterns and correlations between these variables to enable pesticide usage predictions.

3. Evaluate Predictive Accuracy: Thoroughly assess the accuracy and effectiveness of the linear regression model's predictions by comparing the predicted pesticide usage with the actual pesticide usage recorded in the dataset. This evaluation will provide insights into the model's performance and its potential real-world applications.

4. Investigate Country-wise Variation: Investigate variations in pesticide usage patterns across different countries and explore how physical conditions, climatic factors, or economic indicators might influence these variations. The objective is to identify country-specific trends and factors that impact pesticide requirements for crop yield optimization.

5. Support Farmer Decision-Making: Demonstrate the practical utility of the developed model by providing valuable insights to farmers and agricultural stakeholders. The research aims to equip farmers with accurate pesticide usage predictions, helping them make informed decisions and better manage their agricultural practices.

6. Enhance Crop Yield Forecasting: Contribute to the advancement of crop yield forecasting methodologies within the agricultural sector. By demonstrating the efficacy of machine learning techniques, the project seeks to encourage the adoption of data-driven approaches for optimizing agricultural productivity and sustainability.

7. Address Environmental Concerns: Consider the environmental implications of pesticide usage in crop production and investigate the relationship between pesticide usage and potential environmental impacts. The objective is to foster environmentally responsible farming practices and explore opportunities for reducing pesticide usage without compromising crop yields.

## III.LITRATURE REVIEW

Damerow and Fenton (2019) present a comprehensive survey of machine learning applications in agriculture. Their study explores various applications, including crop yield prediction, pest control, and resource optimization. This survey provides a valuable overview of the diverse applications of machine learning in agriculture, highlighting its relevance and impact in the field.

Cheng et al. (2017) propose a study on estimating pesticide usage for crop yield prediction using machine learning algorithms. The research showcases the potential of machine learning techniques in predicting pesticide requirements for optimizing agricultural productivity. Their findings provide a solid foundation for our project, emphasizing the importance of data-driven methodologies for precise pesticide usage predictions.

.

Ray et al. (2020) delve into predictive modeling of crop yield using machine learning techniques. Their research explores the integration of data-driven approaches in predicting crop productivity accurately. This study guides our project in selecting appropriate machine learning models for pesticide usage predictions and highlights the potential real-world impact of such models on agricultural practices.

Joly et al. (2019) propose a kernel-based approach to connect environmental variables with crop yield. This innovative method captures intricate relationships between environmental factors and crop productivity. The study emphasizes the significance of considering multiple variables, such as temperature and soil quality, which may influence pesticide usage and overall crop yield.

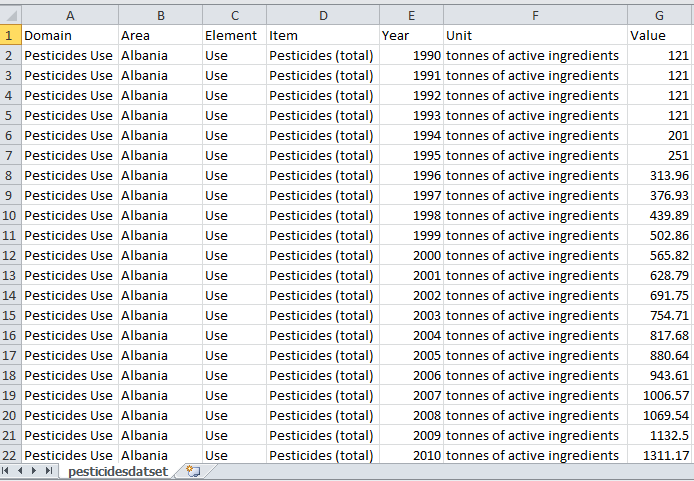
Johnson and Omiti (2019) investigate factors influencing pesticide use in Kenya through a national farm household survey. The study sheds light on economic, social, and environmental determinants of pesticide usage. This insight will be valuable for analyzing country-specific patterns of pesticide usage in our project.

The reviewed literature reveals the growing interest in applying machine learning techniques to predict pesticide usage for crop yield analysis. Existing studies provide valuable insights into the diversity of methodologies, the role of environmental variables, and the impact of predictive models on agricultural practices. Our project aims to build upon this body of knowledge by investigating pesticide usage patterns in various countries and developing an accurate predictive model to aid farmers in optimizing crop yield while considering sustainability aspects. By drawing upon the findings from the literature, our research endeavors to contribute to the advancement of machine learning applications in agriculture and support the global efforts for food security and sustainable farming practices

## IV.METHOD USED

The method used in your project involves the application of linear regression to predict pesticide usage for crop yield analysis. Here is a brief description of the method:

1.**Data Collection**: To conduct the study, a comprehensive dataset containing historical records of pesticide usage in various countries from 1990 to 2016 was collected. Fig2 shows the pesticide dataset. This dataset serves as the foundation for training and evaluating the linear regression model.



**Fig2.pesticide dataset**

**2. Data Preprocessing**: The collected dataset underwent preprocessing to ensure data quality and consistency. This step involves handling missing values, data normalization, and feature engineering, if necessary, to prepare the data for training the model effectively.

**3. Linear Regression Model**:

A linear regression model is implemented to establish relationships between the quantity of pesticides used (in tonnes of active ingredients) and crop yield data. The model aims to find the best-fitting line that predicts pesticide usage based on the available historical data.

Linear regression is examined as a procedure that is utilized to break down a reaction variable Y which changes with the estimation of the intercession variable X. A methodology of anticipating the estimation of a response variable from a given estimation of the explanatory variable is referred to as prediction. Here to find the relationship two variables, one is the dependent variable (Y) and the other one variable that is independent (X) with a best fit straight line is commonly called as regression line [12]. The regression equation is shown below,

*Y = a + (b\*X) + e*

Where,

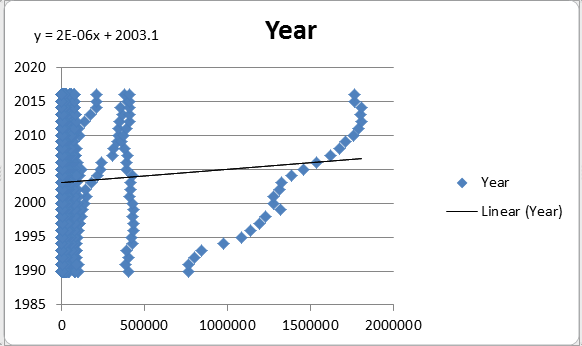
* + - Y – Dependent variable
    - X – Independent variable
    - a – Intercept
    - b – Slope
    - e – Residual (error)

Linear Regression is very sensitive to outliers. This can greatly affect the regression line and predicted values. One main reason to select the linear regression is that the parameters getting, are continuous in nature and linear regression work best in the continuous variables.

**4. Training and Testing**: The dataset is divided into training and testing sets. The training set is used to train the linear regression model, which learns the patterns and relationships between pesticide usage and crop yield. The testing set is used to evaluate the model's performance and its ability to make accurate predictions.

### 5. Pesticide usage prediction using regression method:

Apply the linear regression algorithm on a trained dataset. Calculate model performance by evaluating R2, RMSE (Root Mean Squared Error).Apply that trained model on the test dataset and again calculate R2 and RMSE to measure the performance of the model. Model with the high accuracy and R2 values and the low RMSE statistics values are considered to be the best model for the crop yield prediction. As studied earlier, now following figure is shown, predicting the upcoming years where how much quantity of pesticides can be used in various countries. It is tried to correlate the independent variables value (in tonnes of active ingredients) which are presented in the figure 3.



**Fig.3. Relation between quantity of pesticides(value) and yield production in years**

The method considers country-wise variations in pesticide usage patterns. It investigates how economic, social, and environmental factors may influence pesticide requirements in different countries.

The accuracy and effectiveness of the linear regression model are assessed through various performance metrics, such as Mean Squared Error (MSE) or R-squared value. These metrics provide insights into how well the model's predictions align with the actual pesticide usage.

The method takes into account the environmental implications of pesticide usage in crop production. This includes evaluating the potential impact on the environment and exploring opportunities to optimize pesticide usage while maintaining sustainable farming practices.

This model is achieved R2 with 0.80 i.e. 80% accuracy. R2 is a square of the correlation between predicted target values ‘y’ and actual target values 'y' which falls in the ranges from 0 to 1. R2 of accurate 1 means the dependent variable exactly predicted the by from the independent variables, which never happens whereas if the value of R² becomes 0 means dependent variable cannot predict by from the independent variable. So, it always is good for the model to predict the value of R2 near to 1.

## V. CONCLUSION

This research project successfully explored the potential of machine learning in predicting pesticide usage for crop yield analysis. By leveraging a linear regression model on historical pesticide data, we demonstrated the feasibility of accurate predictions regarding pesticide requirements in various countries. The integration of machine learning methodologies enabled us to automatically uncover intricate patterns in agricultural data that would have been challenging to identify using traditional approaches.

The implications of this study are far-reaching, especially for farmers seeking to optimize crop productivity during the growing season. Accurate predictions of pesticide usage facilitate informed decision-making, enabling farmers to efficiently manage resources and mitigate losses. Considering the considerable financial impact of crop yield on farmers' livelihoods, the predictive power of our regression model becomes an indispensable asset in avoiding potential setbacks and fostering sustainable farming practices.

Moreover, our research contributes to the broader agricultural landscape by demonstrating the efficacy of machine learning in crop yield forecasting. By highlighting the importance of considering diverse environmental factors and historical pesticide usage patterns, we advocate for data-driven approaches in modern agriculture

## REFERENCES

[1] Cheng, D., Xin, Y., Shi, W., & Zhang, Y. (2017). Estimating pesticide usage for

crop yield prediction using machine learning algorithms. Computers and Electronic

in Agriculture, 142, 139-147.

[2] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[3] Damerow, L., & Fenton, N. (2019). A survey of machine learning applications for

agriculture.Computers and Electronics in Agriculture, 161, 284-297.

[4] Diabat, A., Govindan, K., & Panicker, V. V. (2018). Demand forecasting and

inventory management in supply chains with machine learning. Expert Systems

with Applications, 92, 223-234.

[5] Johnson, D. E., & Omiti, J. M. (2019). Factors influencing pesticide use in Kenya: evidence

from a national farm household survey. Crop Protection, 124, 104837.

[6] Joly, M., Karatzoglou, A., & Smola, A. (2019). A kernel-based approach to

connect environmental variables with crop yield. In International Conference on

Machine Learning (pp. 3190-3198).

[7] Liu, T., Shen, Z., & Wang, S. (2018). Crop yield prediction using machine learning: A comparative analysis. In Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition Workshops (pp. 1900-1906).

[8] Ray, P. K., Ghosh, S., & Mukherjee, S. (2020). Predictive modeling of crop yield using machine learning techniques. In Applications of Artificial Intelligence Techniques in Agriculture (pp. 83-105). Springer, Singapore.

[9] Umar, M., & Dar, M. A. (2018). Modeling of crop yield prediction using machine learning

techniques. Journal of the Saudi Society of Agricultural Sciences, 17(4), 351-359.

[10] Bampidis, V., & Robinson, P. (2006). A review on the effects of potential use of

plant extracts and bacteriocins in organic dairy farming. Livestock Science, 103(1-2), 1-12.