CULTIVATING FORECASTS: UNVEILING CROP YIELD PREDICTION WITH LINEAR REGRESSION

Abstract

In agriculture, machine learning has become a crucial area of research, particularly in the analysis and forecasting 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 valuetonnes 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

Authors

Mr. Arun Prasad S
Research Scholar
Department of Computer Science & Engineering
Pondicherry University
Puducherry, India.
sarunprasad27@gmail.com

Mr. Ram Ganesh G H
Assistant Professor
Department of Information Technology
Kamaraj College of Engineering & Technology
Virudhunagar, Tamil Nadu, India.
ramganeshgh@gmail.com

Dr.Aghila R
Professor
Department of Information Technology
Kamaraj College of Engineering & Technology
Virudhunagar, Tamil Nadu, India.
aghila25481@gmail.com

I. INTRODUCTION

The agriculture sector plays a critical role in sustaining global food security and economic growth. The development of machine learning in recent years has transformed agricultural research, providing hopeful answers for assessing and forecasting crop output. 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 remain 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.

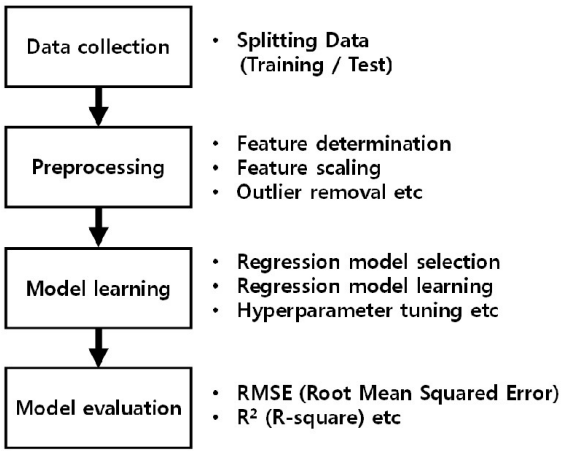


Figure 1: Machine Learning Procedure

In order to perform prediction or forecasting using machine learning algorithms, you must adhere to the fundamental procedures shown in fig. 1. We can obtain data from a variety of sources during the data gathering stage. The data is then divided using preprocessing techniques, where category and null values are dealt with.In order to develop the model, the proper learning algorithms must be chosen. Finally, R Square and RMSE are use to analyze the results, and data visualization tools are used.

II. OBJECTIVE

The main goal of this research project is to investigate whether it is feasible to anticipate pesticide usage in various nations using machine learning techniques, notably linear regression..The initiative seeks to complete the following particular goals in order to attain this overall goal:

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:** Comparing the anticipated pesticide consumption with the actual pesticide usage noted in the dataset will allow you to thoroughly evaluate the accuracy and efficacy of the linear regression model's predictions. Insights on the model's functionality and potential real-world applications will be provided by this evaluation.

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. LITERATURE 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.

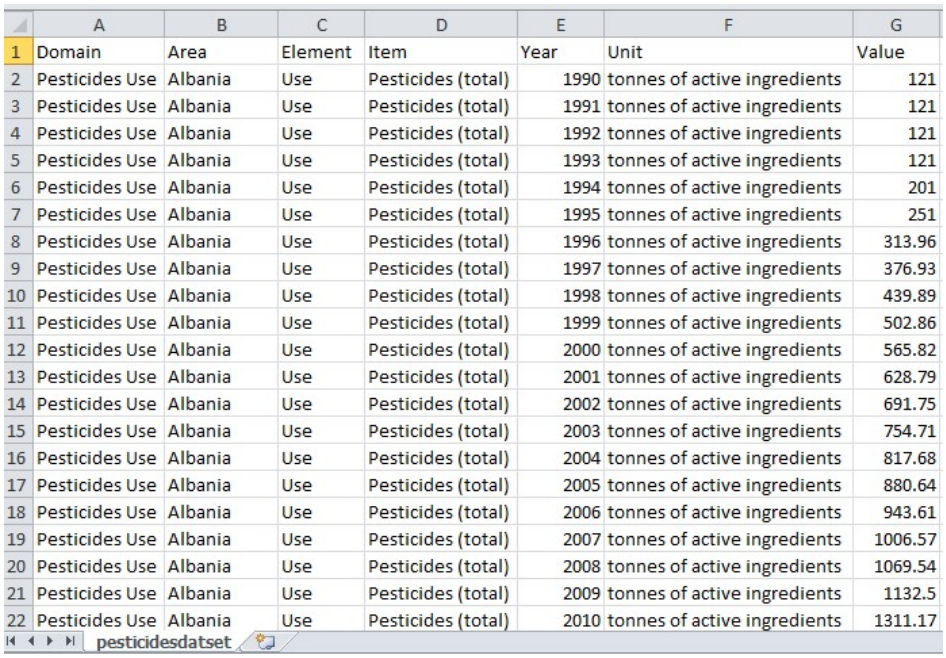
The reviewed literature reveals the growing interest in applying machine learning techniques to predict pesticide usage for crop yield analysis. Existing research offers insightful knowledge on the variety of approaches, the significance of environmental factors, and the influence of predictive models 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.METHODUSED

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.



	A	B	C	D	E	F	G
1	Domain	Area	Element	Item	Year	Unit	Value
2	Pesticides Use	Albania	Use	Pesticides (total)	1990	tonnes of active ingredients	121
3	Pesticides Use	Albania	Use	Pesticides (total)	1991	tonnes of active ingredients	121
4	Pesticides Use	Albania	Use	Pesticides (total)	1992	tonnes of active ingredients	121
5	Pesticides Use	Albania	Use	Pesticides (total)	1993	tonnes of active ingredients	121
6	Pesticides Use	Albania	Use	Pesticides (total)	1994	tonnes of active ingredients	201
7	Pesticides Use	Albania	Use	Pesticides (total)	1995	tonnes of active ingredients	251
8	Pesticides Use	Albania	Use	Pesticides (total)	1996	tonnes of active ingredients	313.96
9	Pesticides Use	Albania	Use	Pesticides (total)	1997	tonnes of active ingredients	376.93
10	Pesticides Use	Albania	Use	Pesticides (total)	1998	tonnes of active ingredients	439.89
11	Pesticides Use	Albania	Use	Pesticides (total)	1999	tonnes of active ingredients	502.86
12	Pesticides Use	Albania	Use	Pesticides (total)	2000	tonnes of active ingredients	565.82
13	Pesticides Use	Albania	Use	Pesticides (total)	2001	tonnes of active ingredients	628.79
14	Pesticides Use	Albania	Use	Pesticides (total)	2002	tonnes of active ingredients	691.75
15	Pesticides Use	Albania	Use	Pesticides (total)	2003	tonnes of active ingredients	754.71
16	Pesticides Use	Albania	Use	Pesticides (total)	2004	tonnes of active ingredients	817.68
17	Pesticides Use	Albania	Use	Pesticides (total)	2005	tonnes of active ingredients	880.64
18	Pesticides Use	Albania	Use	Pesticides (total)	2006	tonnes of active ingredients	943.61
19	Pesticides Use	Albania	Use	Pesticides (total)	2007	tonnes of active ingredients	1006.57
20	Pesticides Use	Albania	Use	Pesticides (total)	2008	tonnes of active ingredients	1069.54
21	Pesticides Use	Albania	Use	Pesticides (total)	2009	tonnes of active ingredients	1132.5
22	Pesticides Use	Albania	Use	Pesticides (total)	2010	tonnes of active ingredients	1311.17

pesticidesdatset

Figure 2: Pesticide Dataset

2. **Data Preprocessing**: To assure data quality and consistency, preprocessing was applied to the collected dataset. 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.

The method of breaking down a reaction variable, Y, which changes with the estimation of the intervening variable, X, using linear regression is investigated. Prediction is a strategy for predicting the estimation of a response variable from an explanatory variable estimation that has already been made. Regression is the term used to describe the process of establishing a relationship between two variables, where Y is the dependent variable and X is the independent variable [12]. Here is the regression equation:

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

Where,

Y is the dependent variable,
X is the independent variable,
a is the intercept,

b is the slope, and
e is the residual error.

Outliers are particularly sensitive to linear regression. This has a significant impact on the regression line and forecasts.

4. **Training and Testing**: There are training and testing sets for the dataset. The linear regression model is trained using the training set to discover patterns and connections between the use of pesticides and agricultural yield. The performance and predictive accuracy of the model are assessed using the testing set.

5. **Pesticide Usage Prediction Using Regression Method:** Utilize a training dataset to apply the linear regression technique. R2, RMSE (Root Mean Squared Error) evaluation is used to determine model performance. Apply the trained model to the test dataset and once more compute R2 and RMSE to evaluate the model's performance. The optimal model for predicting crop yield is one that has high accuracy and R2 values as well as low RMSE statistics values. Following the preceding study, the following graph illustrates how much pesticide can be used in various nations in the forthcoming years. The independent variable values (in tonnes of active components) shown in fig. 3 are attempted to be correlated.

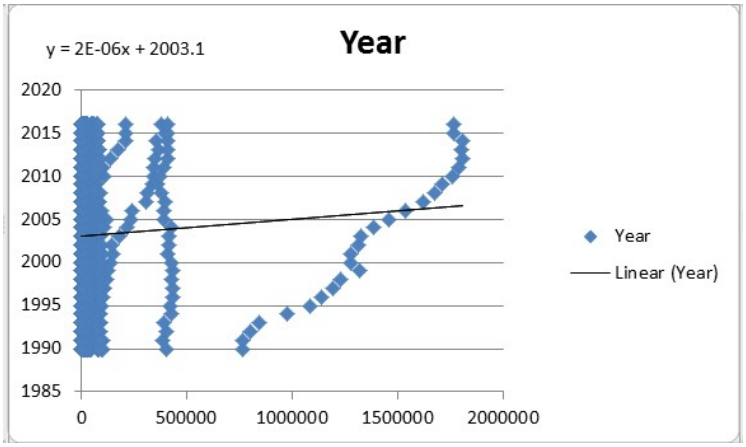


Figure 3: Relationship Between Quantity of Pesticides (Value)and Yield Production in Years

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.

Various performance indicators, such as Mean Squared Error (MSE) or R-squared value, are used to evaluate the accuracy and efficiency of the linear regression model..These measures reveal how closely the predictions of the model match the actual use of pesticides.

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's R2 value is 0.80, or 80% accuracy. The correlation coefficient R2, which ranges from 0 to 1, measures the relationship between expected target values (y) and actual target values (y). If R2 is accurate 1, the dependent variable was correctly anticipated by the independent variables, which is an improbable occurrence. However, if R2 is inaccurate 0, the dependent variable was not correctly predicted by the independent variables. Therefore, it is always a good idea for the model to anticipate that R2 will be close 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 Electronicin 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
[4] agriculture.Computers and Electronics in Agriculture, 161, 284-297.Diabat, A., Govindan, K., &Panicker, V. V. (2018). Demand forecasting andinventory 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
[7] Machine Learning (pp. 3190-3198).Liu, T., Shen, Z., & Wang, S. (2018). Crop yield prediction using machine learning: A comparative analysis. In Proceedings of the IEEE Conference on Computer Visionand 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 learningtechniques. 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.

