Research Article

Dataset Analysis and Feature Characteristics to Predict Rice Production based on eXtreme Gradient Boosting

Ella Budi Wijayanti[1], De Rosal Ignatius Moses Setiadi[1,*], and Bimo Haryo Setyoko[2]

[1] Faculty of Computer Science, Dian Nuswantoro University, Semarang, Central Java 50131, Indonesia; email: ellaabw@gmail.com, moses@dsn.dinus.ac.id
[2] Pusat Teknologi Informasi dan Pangkalan Data, Universitas Islam Negeri Salatiga, Indonesia; e-mail : bimo_hs@uinsalatiga.ac.id
* Corresponding Author : De Rosal Ignatius Moses Setiadi

Abstract: Rice plays a vital role as the main food source for almost half of the global population, contributing more than 21% of the total calories humans need. Production predictions are important for determining import-export policies. This research proposes the XGBoost method to predict rice harvests globally using FAO and World Bank datasets. Feature analysis, removal of duplicate data, and parameter tuning were carried out to support the performance of the XGBoost method. The results showed excellent performance based on R^2 which reached 0.99. Evaluation of model performance using metrics such as R^2, MSE, and MAE measured by k-fold validation show that XGBoost has a high ability to predict crop yields accurately compared to other regression methods such as Random Forest (RF), Gradient Boost (GB), Bagging Regressor (BR) and K-Nearest Neighbor (KNN). Apart from that, an ablation study was also carried out by comparing the performance of each model with various features and state-of-the-art. The results prove the superiority of the proposed XGBoost method. Where results are consistent, and performance is better, this model can effectively support agricultural sustainability, especially rice production.

Keywords: Harvest prediction; Paddy production forecasting; Regression analysis; Rice yield prediction; Rice production prediction; XGBoost Prediction.

Received: January, 24th 2024
Revised: February, 20th 2024
Accepted: February, 21th 2024
Published: February, 21th 2024

1. Introduction

Rice is an essential commodity because it acts as a food source for about half of the world's population[1]. On a global scale, rice covers more than 21% of human calorie needs and is important in supporting food security and community welfare. Global rice production in 2021 will reach 787 million tons. Therefore, a rapid transition towards a sustainable food and agricultural system is needed to maintain global food security, specifically in this case rice[2]–[4]. Predicting rice production is one way to achieve this, because accurate predictions can determine the right policy[2].

Predictions of crop production are highly dependent on weather conditions, pesticide use, and rice yield data per hectare [3]. All these features are very important in making decisions regarding correlated features, if the features are not correlated with each other, this can have a significant impact on the overall performance of the prediction results[4]. The overall prediction results aim to identify complex patterns using data from various sources, to increase the accuracy of rice yield predictions. This approach is crucial in understanding the dynamics of relationships between variables and supporting better decision making in the global agricultural context[5]. Technological developments in various sectors, including agriculture, make it possible to use it to predict more accurate rice yields. Inaccurate predictions can hinder efficiency in agricultural planning and resource management[6]. By using historical data, predictive technology in data mining can help identify relationship patterns that influence crop yields. Identifying this requires supporting features[7].

Prediction of rice harvest yields can be made using machine learning (ML) methods, for example, Random Forest (RF)[8], [9]; XGBoost[10]–[12]; Gradient Boosting(GR)[13], [14];

DOI : 10.62411/jcta.10057

K-Nearest Neighbor (KNN)[15], [16]; and Bagging Regressor (BR)[17], [18]. The advantage of RF is that it utilizes a large number of decision trees to overcome overfitting, but the disadvantage is that it is less efficient in handling complex non-linear relationships and can be difficult to interpret[19]. GR effectively handles data complexity and is robust against overfitting but may require longer computing time. KNN has the advantage of handling complex patterns but is susceptible to outliers and requires precise parameter settings [20]. BR effectively reduces model variance but may be less efficient in handling complex data [17], [18]. XGBoost is one of the most influential and high-performance machine learning techniques. It is used to predict rice yields. Its strength is its ability to handle a variety of complex features and problems in data analysis and predictive model building. However, its drawback is that it tends to overfit, which means it can overfit the training data and is difficult to interpret directly[21]. But by using the right combination of features, XGBoost can be a powerful choice for solving complex prediction tasks[22].

Research [23] compared the RF and XGBoost algorithms and proved that XGBoost has an accuracy of 84.79%, which is better than RF, which is only 82.48%. Another study [24] also compared XGBoost, Gradient boosting, Bagging regression, LR. The result is that the XGBoost method is also superior to other methods, with an accuracy of 98.1%. Another advantage of XGBoost is its higher execution speed and good model performance, especially on well-structured or tabular datasets[25]. However, the XGBoost method will not be optimal without using the right features. Feature selection has a significant impact on prediction results in the field of data mining[26], [27]. Therefore, it is important to choose features carefully. However, suppose a dataset has been designed with knowledge of the features with the highest influence. In that case, this will improve the dataset's quality and produce the best performance when applied with various algorithms[27]. In some production prediction research, credible public datasets can be used as datasets. FAO and World Bank. These two datasets have been combined in research[5], [28], and the results are better prediction performance. This proves that the two datasets are correlated to support better decision-making in the context of global agriculture.

Predictions can be evaluated with several metrics, namely, the coefficient of determination (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE)[29]. R^2 is a measure that indicates the extent to which the regression model is able to explain variations in the data, with values ranging between 0 and 1[6]. The MAE metric measures the average absolute difference between predicted and actual values does not give more weight to large errors, and is resistant to outliers, while MSE is a metric that measures the extent of the model's prediction error in the same units as the dependent variable[2]. Using various measuring instruments will provide a more in-depth analysis of the results. Based on the literature above, this research aims to analyze the correlation between features further and measure the XGBoost method's prediction performance. In more detail, the objectives of this research are:

1. Analyze the influence of dataset features on prediction performance.
2. Tuning XGBoost by setting parameters to get the best performance.
3. Analyze in more detail the prediction results based on R^2, MSE, MAE.
4. Comparing several other ML models and XGBoost to prove XGBoost's superiority.

The remainder of this article is presented in three parts: methodology, the second part discussing theory, literature, methods used, and reasons. The third section explains the results and analysis; the last is the conclusion.

2. Related Works

Several previous studies have conducted research regarding rice harvest prediction using ML. Ge et al. [24] tested the XGBoost method and compared it with several ML methods, such as support vector regression (SVR), Linear Regression (LR), KNN, RF, BR, GB, and AdaBoost. The result is that XGBoost is the best method because it shows a higher R^2 level in predicting crop yields of 93.91% compared to other algorithms tested.

Singha and Swain [3]researched rice yield predictions, specifically using the FAO dataset with the World Bank. The research results show that applying the RF method produces a positive correlation in predicting rice yields. Evaluation of regression metrics shows that the R^2 level reaches 86%, with an MAE of 0.88 and an MSE of 1.23. Additionally, Cedric et al. [28] studied crop yield prediction using combines climate, weather and agricultural yield data, from FAO and World Bank datasets. The findings show a positive correlation in this context by applying the KNN method. In the evaluation of regression metrics, this model achieved an R^2 level of 95.03% and an MAE of around 0.160.

Based on several studies that have been carried out on crop yield prediction objects using the ML approach, it appears that the XGBoost method gets the best results. However, these three studies have different datasets. The FAO and World Bank datasets can be used as references because they are valid and widely used in other research. The XGBoost method, considered superior, needs to be tested with several other methods, such as RF, KNN, and several other ML methods. The final results obtained from this research will be evaluated using several measuring tools, such as R^2, MAE, and MSE, and compared with previous research.

3. Research Methodology

This section begins by explaining the research methodology presented in Figure 1. The methodology used in this research is simple, consisting of four main stages: dataset collection, preprocessing, XGBoost prediction, and evaluation, presented in subchapters 3.1 to 3.4.

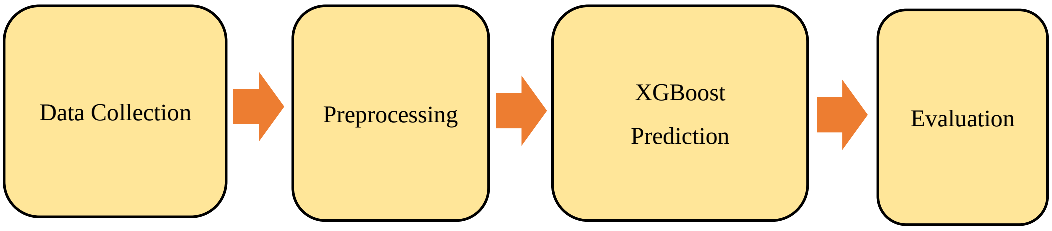


Figure 1. Research Stages.

3.1. Dataset Collection

At this stage, global agricultural rice harvest sector data is collected from the Food and Agriculture Organization (FAO), which includes information regarding harvest yields and pesticide use, which is downloaded from the link https://www.fao.org/faostat/en/#data/QCL as well as from the World Bank which provides data related to rainfall and temperature which is downloaded from the link https://databank.worldbank.org/source/world-development-indicators/ or to make easier we uploaded at GitHub as mirror URL https://github.com/ellabw/Crop-Yield-Prediction. Each dataset is equipped with features, as in Table 1 for the FAO dataset and Table 2 for the World Bank dataset.

Table 1. Dataset FAO.

No	Attribute	Data Type	Note
1.	Area	Object	Country/Area/Region
2.	Item	Object	Item name (Rice or Paddy)
3.	hg/ha_yield	Int	Production yield in hectograms per hectare (Hg/Ha)
4.	pesticides_tonnes	Float	Pesticides used (tons)

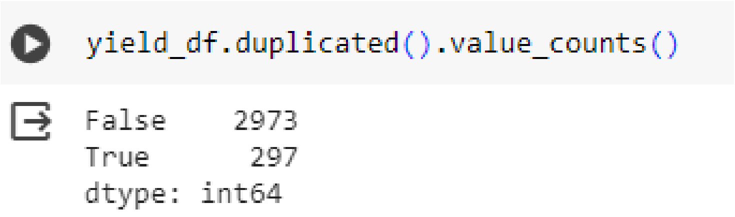
Table 2. Dataset World Bank.

No	Attribute	Data Type	Note
1.	average_rain_fall_mm_per_year	Float	Average rainfall per year
2.	avg_temp	Float	Average temperature

Based on the data presented above, the FAO dataset has four main features, namely information on crop yields and pesticide use, which have object and integer values, and the World Bank has two features, namely rainfall and temperature, which have float values. These two datasets are directly related and correlated, allowing them to produce the best performance on the dataset. The reason for using a combination of two datasets is that it forms a more comprehensive feature set, which is then integrated into the machine learning model. The aim is to improve performance and accuracy in analyzing and predicting crop yields [28].

3.2. Preprocessing

This stage helps identify comprehensive patterns, supports better decision-making, improves performance, and forms the basis for advanced analysis. After combining, the total data becomes 3,270 with six columns and is stored in the yield_df DataFrame. The following preprocessing step involves removing duplicate data to clean the dataset and ensure accuracy and speed in the data mining process, removing duplicates using the duplicate() function in the pandas' library in Python. Here, the duplicated() function returns a Boolean Series that marks each row in the yield_df DataFrame as True if the row duplicates the previous row and False if not. The value_counts() function is then used to count the number of True and False values produced by duplicated(), see Figure 2. In the merged dataset, 297 duplicate data were identified, leaving 2973 data after deletion, as seen in Figure 3. This aims to improve the performance and efficiency of data analysis.



```
yield_df.duplicated().value_counts()
```

```
False    2973
True      297
dtype: int64
```

Figure 2. Removing duplicates and its functions.

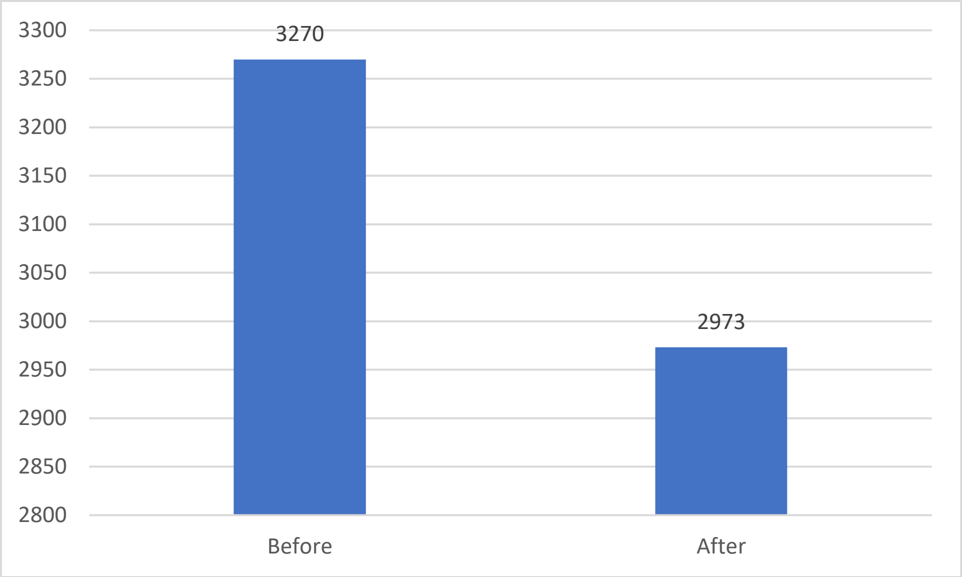


Figure 3. Total record dataset before and after deleting duplicate data.

The next step in data analysis is to identify the presence of empty values. In this test, using the yield_df.isnull().sum() function, the results show that no empty values were found in the dataset after the previous preprocessing process.

Then, the next preprocessing step is to remove unused attributes. The combined dataset initially had six attributes, but the "item" attribute was removed because it only included one object, namely rice. Information attributes about historical crop yields, such as area and production value in hectograms per hectare (Hg/Ha) are retained because these variables are considered independent variables that can help the regression model understand historical patterns of crop yields. Attributes related to weather conditions, such as rainfall and average temperature, are also retained because, in regression analysis, weather conditions become predictor variables that directly impact crop yields. In addition, the attribute of pesticide use is also maintained because it can provide an understanding of its impact on crop yields. The dataset can be analyzed to improve prediction performance by retaining these features in the regression model.

3.3. Prediction Method

This research proposes XGBoost as a method for predicting rice yields. XGBoost has several important parameters, namely n_estimators, random_state, learning_rate. If the n_estimators value is high, the model complexity increases, and vice versa, if it is low, the model complexity decreases. While a low learning_rate will slow down learning, affecting the trade-off between model complexity and generalization, vice versa, it can speed up learning, affecting the trade-off between model complexity and generalization and increasing the risk of overfitting. At the same time, the random_state parameter is used to control randomness in tree generation, ensuring consistent results for repeatable experiments. The value, which can be from 0 to infinity, does not affect the model's performance directly but allows the reproduction of the same results using the same value on re-execution. Table 3 shows the default and after XGBoost tuned parameter values.

Table 3. XGBoost Hyperparameter Setup.

Parameter	Defaults Values	Tuned Values
n_estimators	100	150
learning_rate	0.3	0.1
random_state	42	42

In this research, these parameter values can optimize the results and obtain an optimal model. This value was obtained from several experiments, and the optimal parameters were chosen because they balance good and efficient performance. It is important to note that XGBoost has the advantage of handling non-linear dependencies and can provide good results for regression problems such as rice yield prediction. XGBoost can be a powerful choice for crop yield prediction tasks by understanding the characteristics of supporting feature datasets.

3.4. Evaluation

In this research, the evaluation of prediction methods was carried out using several metrics, such as R^2, MAE, and MSE. Where R^2 is used to measure how well the regression model fits observational data. MAE measures the average of the absolute differences between predicted and observational values, which gives an idea of how big the overall prediction error is. MSE measures the average of the squared differences between predicted and observational values, emphasizing large errors, as errors are squared before being calculated. So, when using regression methods, evaluating model performance often involves metrics such as R^2, MAE, and MSE as measuring tools to measure the extent to which the model can provide accurate predictions.

4. Results and Analysis

Feature analysis is carried out before making predictions. Here, a data visualization presents the complexity of the relationships between variables in a dataset. One of the visualizers used is the correlation heatmap, which is presented in Figure 4. Based on this figure, several significant findings regarding the relationship between features in the agricultural dataset were found. The strongest negative correlation is between rice yield and average temperature (avg_temp), with a value of -0.68, indicating that increasing temperature is correlated with decreasing crop yield. These findings have important implications in the context of climate change. Furthermore, the moderate negative correlation between area size and pesticide use draws attention to the efficiency of pesticide use over larger areas or different farming practices. However, the very low correlation between pesticide use and crop yield (value -0.044) shows the complexity of this relationship, so this study also performed tests by eliminating the pesticide features presented in Table 6.

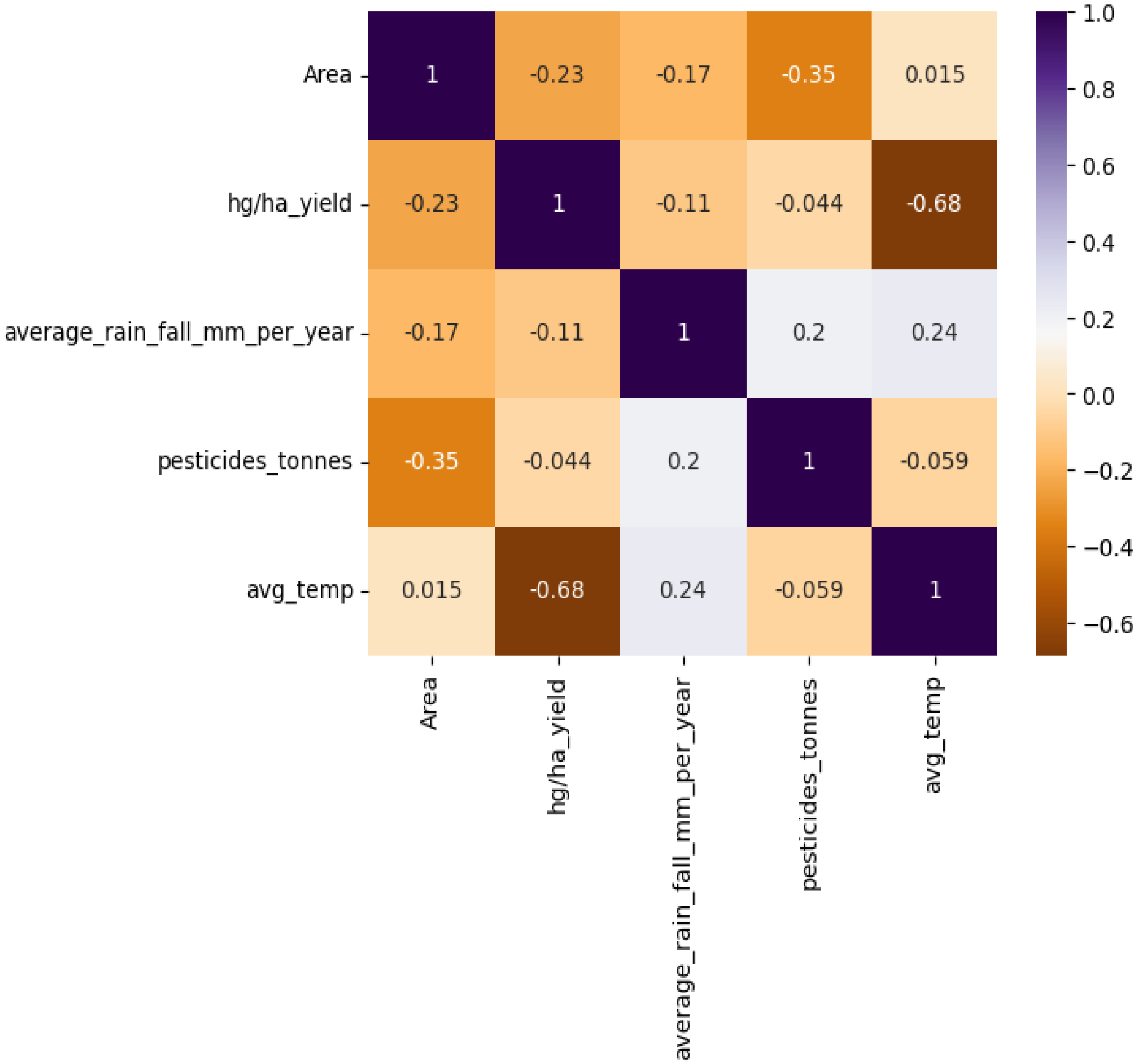


Figure 4. Heatmap of correlation between features.

Figure 5 explains that the pair plot is a statistical tool used to explore the relationship between variables in a dataset. Each variable has a histogram/density plot, providing an overview of the data distribution. Scatterplots show relationships between pairs of variables, enabling analysis of patterns, trends, and identification of outliers. Colors and legends are used to differentiate country data, while correlation can be evaluated by assessing the direction and trend of data points. Pair plots are very useful for testing the influence of variables, such as average temperature, on crop yields. This analysis can reveal the consistency of relationships across countries or variations based on regional context. This graphic matrix provides in-depth insight into the relationship of variables in agriculture and provides a holistic understanding of the complex relationships of key agricultural variables.

Overall, the three graphs in Figure 5 provide an in-depth understanding of the complex interactions between variables in the agricultural context. The first graph is a Density Plot for hg/ha_yield, visually representing the distribution of yields per hectare with peaks identifying the most common values in the dataset. Correspondingly, the second graph, namely the Scatterplot between rainfall and hg/ha_yield, provides insight into the relationship between annual rainfall and crop yields. By looking at the distribution pattern of the points, we can find out whether there is a certain trend or pattern that indicates a relationship between the two variables. Meanwhile, the scatterplot graph between avg_temp and hg/ha_yield, allows us to evaluate the relationship between average temperature and crop productivity.

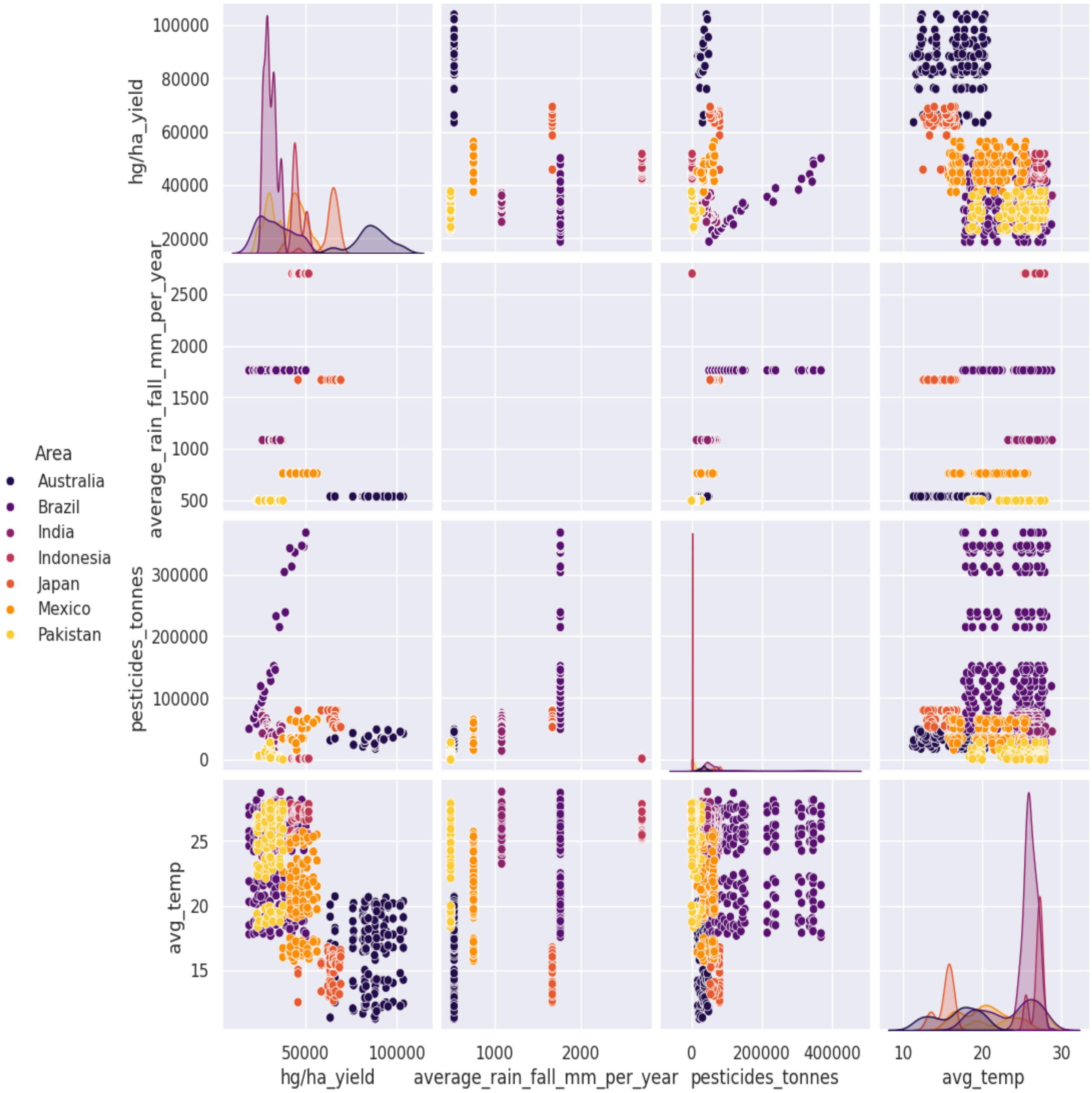


Figure 5. Pair plot between variables in the dataset.

Figure 6 shows three scatterplots. Scatterplot analysis comparing rice yields with three variables (pesticide use, annual rainfall, and average temperature) provides insight into the factors influencing crop yields. The top graph indicates the relationship between pesticide use and crop yields, while the middle graph shows the impact of annual rainfall. The third graph highlights the relationship between average temperature and crop yield. This analysis allows the identification of patterns, trends, and outliers that provide insight into the agricultural practices and climate conditions that influence rice production in different countries. Predictive models like XGBoost can leverage this information to improve crop yield predictions and support better agricultural decision-making.

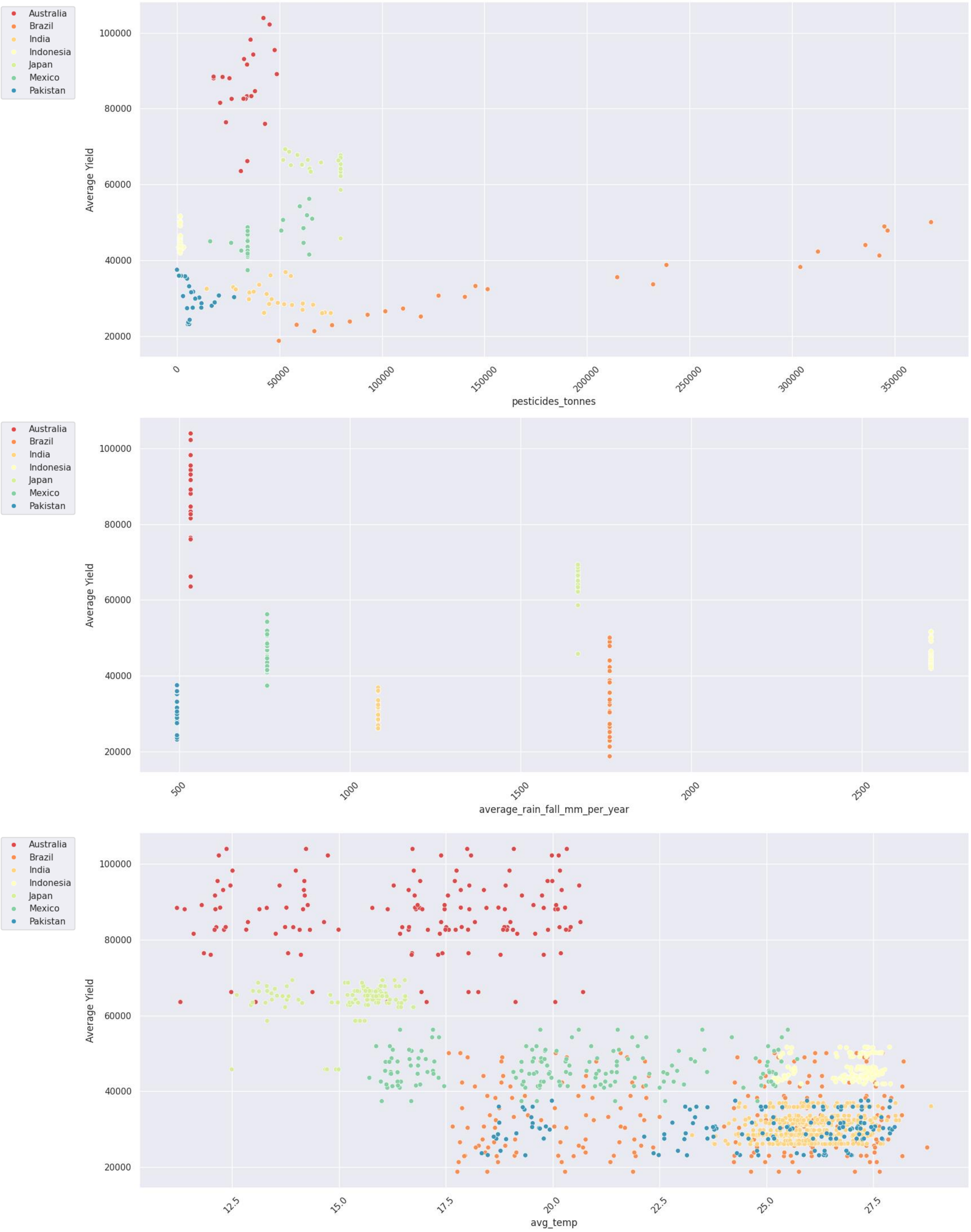
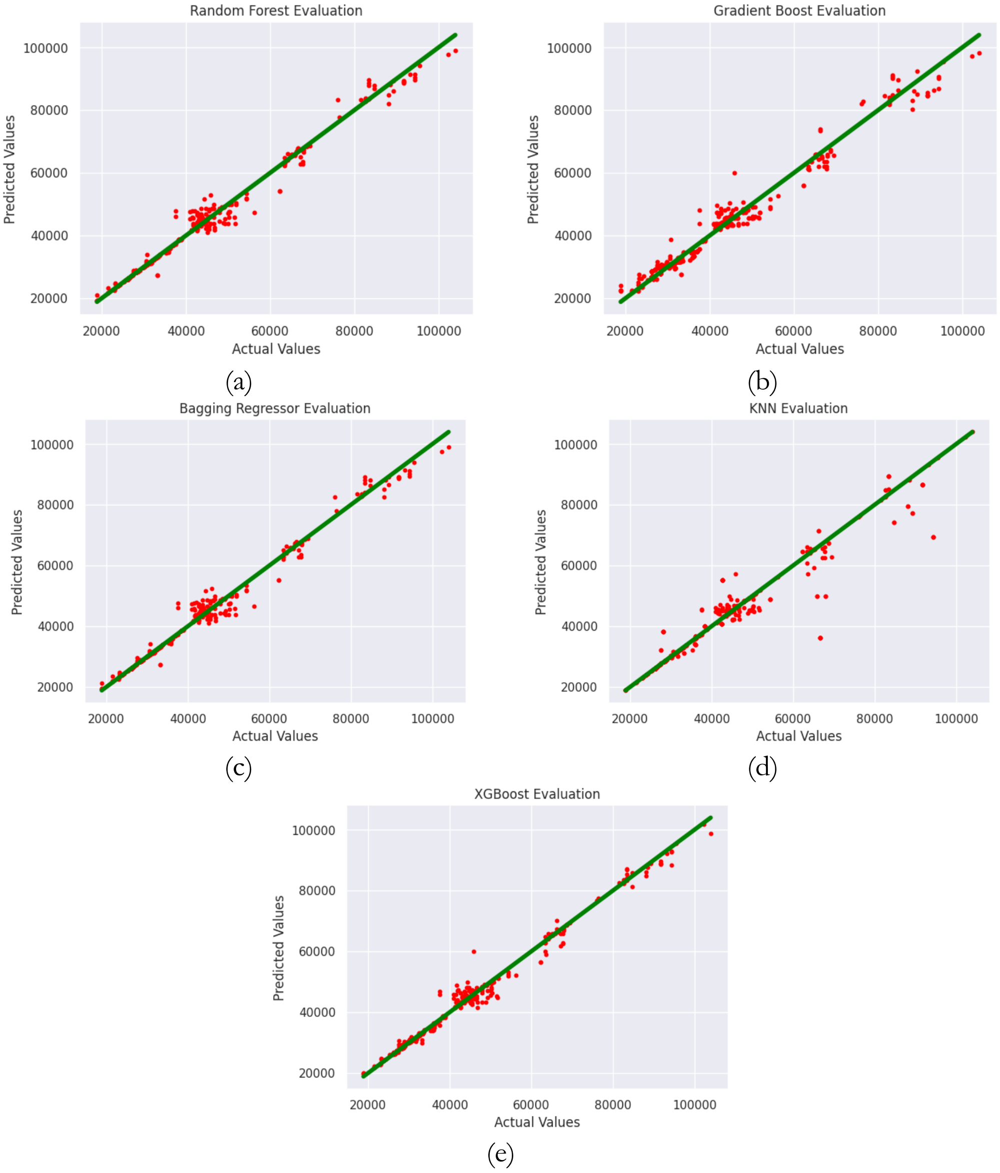


Figure 6. A scatter plot comparing three variables.

Next, XGBoost is implemented in Python using the sci-kit-learn library and the Extreme Gradient Boosting class for machine-learning regression tasks. Several other regression models were also compared, such as RF, GR, KNN, and BR. For each model, the code fits the model to the training data (X_train, y_train), performing predictions of the target variable on the test data (X_test). Then, the performance results for the training and test sets are visualized via scatter plots, which depict the predicted values of each model against the actual values, see Figure 7.


Random Forest Evaluation

(a)

Gradient Boost Evaluation

(b)

Bagging Regressor Evaluation

(c)

KNN Evaluation

(d)

XGBoost Evaluation

(e)

Figure 7. Actual vs. predicted plot evaluation (**a**) Random Forrest; (**b**) Gradient Boost; (**c**) Bagging Regressor; (**d**) KNN (**e**) XGBoost.

The plot referred to in Figure 7 above is useful for identifying patterns, trends, or relationships in data and can provide an intuitive visual depiction of the distribution of variables. The scatter plot results illustrate the comparison between the actual value (x-axis) and the value predicted by the model (y-axis). The green line parallel to the diagonal from bottom left to top right represents the ideal situation where the predictions match entirely the actual values. The red dots indicate the actual predicted locations, with the dot spread illustrating how well the model predicts the corresponding values. The denser and closer to the green line, the better the quality of the model's predictions. Visually, it appears that the BR and XGBoost graphs are the best, but there is a red dot on the XGBoost graph, which is relatively far from the green line. Meanwhile, KNN has many red points that are far from the green line. MSE, MAE, and R2 measurements are presented in Table 4 to obtain more valid results. Tables 5 and 6 also present ablation studies and analysis to compare prediction performance with different features.

Table 4. Evaluate results with 5-fold on FAO and World Bank datasets.

Method	R^2	MAE	MSE
Random Forest	0.987736	10.1486	39.6706
Gradient Boost	0.976593	19.5541	75.7154
KNN	0.920631	19.1538	256.7382
Bagging Regressor	0.987122	10.2638	41.6577
XGBoost (default)	0.989290	10.4032	34.6446
XGBoost (tuned)	**0.991311**	**9.6588**	**28.1083**

Table 5. Evaluate results with 5-fold only on the FAO dataset

Method	R^2	MAE	MSE
Random Forest	0.860413	48.3747	54.2625
Gradient Boost	0.714059	82.4857	111.1555
KNN	0.380729	108.9421	240.7326
Bagging Regressor	0.860442	48.2108	54.2512
XGBoost (default)	**0.871240**	**47.3097**	**50.0535**
XGBoost (tuned)	0.870792	48.5412	50.2279

Table 6. Evaluate the results with 5-fold on the FAO, World Bank dataset and removing the pesticide feature

Method	R^2	MAE	MSE
Random Forest	0.891876	46.3754	39.1111
Gradient Boost	**0.913566**	**40.7427**	**31.2653**
KNN	0.901557	42.9418	35.6091
Bagging Regressor	0.892135	46.3711	39.0172
XGBoost (default)	0.900536	42.6458	35.9784
XGBoost (tuned)	0.910596	41.0048	32.3396

Tables 4, 5, and 6 present the results of a 5-fold cross-validation evaluation of the performance of several regression models using a combination of FAO and World Bank datasets, FAO datasets only, and datasets that remove pesticide features. Evaluation with K-Fold Cross Validation creates more robust measurements to provide a complete picture of model performance on various data subsets. It can be observed that the model performance evaluation in Table 4 shows the best results. This proves that using appropriate features can produce more optimal R^2, MAE, and MSE values. The XGBoost model gets the best results, and all methods simultaneously perform better when using a combination of FAO and World Bank datasets.

Contrasting results are shown in Table 5, where all method performance decreases when using features on the FAO dataset only. This shows that the features in the FAO dataset alone are not enough to provide richer and deeper information so the prediction results are less than optimal. The combination of FAO and World Bank datasets by removing the pesticide feature, which was considered to have an insignificant correlation, actually reduced prediction performance. Even though the correlation with the pesticide feature is the lowest, pesticides play a role in protecting plants from pests and disease, so the unavailability of pesticides can reduce the quality of the harvest. This is proven by the prediction results that are less than optimal without the pesticide feature.

The tuned XGBoost model generally performs best on a combination of all features. Even when implemented on the FAO dataset, tuned XGBoost did not perform better when compared to the default XGBoost. Meanwhile, for the third experiment (without the pesticide feature), the GR was slightly superior to the tuned XGBoost. Furthermore, Table 7 also compares the results with several related studies that combined FAO and World Bank datasets.

Table 7. Comparison with state-of-the-art using FAO and World Bank datasets

Method	R^2	MAE	MSE
Method [3] RF	0.86	0.88	1.23
Method [28] KNN	0.95	0.160	-
Proposed XGBoost	0.99	9.6588	28.1083

Based on Table 7, it appears that the proposed model shows the best performance with an R^2 value of 0.99. However, the MAE and MSE values also tend to be large. A high R^2 value indicates the model explains data variations very well, but large MAE and MSE values indicate significant errors in some individual predictions. This means the model is effective overall but may not be accurate for certain cases. The advantage is that the model is very suitable for general trends in the data, although it may be less precise in specific predictions. This may be influenced by the scale of the data used and the characteristics of the dataset. The presence of a large scale on certain targets or features can cause MAE and MSE values to be high.

5. Conclusions

This research presents using XGBoost for rice yield prediction, showing superior performance compared to other methods with evaluation using R2, MAE, and MSE. The model proves its reliability through K-Fold cross-validation and feature analysis, supporting global agriculture and food security decisions. The combination of the FAO and World Bank datasets can also improve prediction performance, as evidenced by better prediction results compared to using the FAO dataset alone. Feature correlation analysis using heatmaps also does not always show that features with the lowest correlation (pesticides) are better removed. The pesticide feature turns out to have a significant effect on predictions because it plays a role in protecting plants from pests and disease. Unfortunately, the proposed XGBoost model has a high MAE and MSE, so although the proposed model generally produces good predictions, specific predictions may be inaccurate. In the future, this model will certainly need to be developed further to obtain more optimal prediction performance.

Author Contributions: Conceptualization: E.B.W. and D.R.I.M.S.; methodology: E.B.W. and D.R.I.M.S.; software: E.B.W.; validation: E.B.W., D.R.I.M.S., and B.H.S.; formal analysis: E.B.W. and D.R.I.M.S.; investigation: E.B.W. and D.R.I.M.S.; resources: E.B.W.; writing—original draft preparation: E.B.W.; writing—review and editing: D.R.I.M.S., and B.H.S.; visualization: X.X.; supervision: D.R.I.M.S and B.H.S.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] A. R. Muslikh, D. R. I. M. Setiadi, and A. A. Ojugo, "Rice Disease Recognition using Transfer Learning Xception Convolutional Neural Network," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1535–1540, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.1529.

[2] A. Morales and F. J. Villalobos, "Using machine learning for crop yield prediction in the past or the future," *Front. Plant Sci.*, vol. 14, p. 1128388, Nov. 2023, doi: 10.3389/fpls.2023.1128388.

[3] C. Singha and K. C. Swain, "Rice crop growth monitoring with sentinel 1 SAR data using machine learning models in google earth engine cloud," *Remote Sens. Appl. Soc. Environ.*, vol. 32, p. 101029, Nov. 2023, doi: 10.1016/j.rsase.2023.101029.

[4] D. Leising, J. Burger, J. Zimmermann, M. Bäckström, J. R. Oltmanns, and B. S. Connelly, "Why do items correlate with one another? A conceptual analysis with relevance for general factors and network models," PsyArXiv, Jan. 2020. doi: 10.31234/osf.io/7c895.

[5] N. Sansika, R. Sandumini, C. Kariyawasam, T. Bandara, K. Wisenthige, and R. Jayathilaka, "Impact of economic globalisation on value-added agriculture, globally," *PLoS One*, vol. 18, no. 7, p. e0289128, Jan. 2023, doi: 10.1371/journal.pone.0289128.

[6] H. T. Pham, J. Awange, M. Kuhn, B. Van Nguyen, and L. K. Bui, "Enhancing Crop Yield Prediction Utilizing Machine Learning on Satellite-Based Vegetation Health Indices," *Sensors*, vol. 22, no. 3, p. 719, Nov. 2022, doi: 10.3390/s22030719.

[7] S. Brice and H. Almond, "Health Professional Digital Capabilities Frameworks: A Scoping Review," *J. Multidiscip. Healthc.*, vol. Volume 13, pp. 1375–1390, Jan. 2020, doi: 10.2147/JMDH.S269412.

[8] X. Gao, J. Wen, and C. Zhang, "An Improved Random Forest Algorithm for Predicting Employee Turnover," *Math. Probl. Eng.*, vol. 2019, pp. 1–12, Jan. 2019, doi: 10.1155/2019/4140707.

[9] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Comput. Sci.*, vol. 162, pp. 503–513, Jan. 2019, doi: 10.1016/j.procs.2019.12.017.

[10] W. Li, Y. Yin, X. Quan, and H. Zhang, "Gene Expression Value Prediction Based on XGBoost Algorithm," *Front. Genet.*, vol. 10, p. 1077, Jan. 2019, doi: 10.3389/fgene.2019.01077.

[11] G. Abdurrahman and M. Sintawati, "Implementation of xgboost for classification of parkinson's disease," *J. Phys. Conf. Ser.*, vol. 1538, no. 1, p. 12024, Jan. 2020, doi: 10.1088/1742-6596/1538/1/012024.

[12] Q. Zhou and A. Ismaeel, "Integration of maximum crop response with machine learning regression model to timely estimate crop yield," *Geo-spatial Inf. Sci.*, vol. 24, no. 3, pp. 474–483, Nov. 2021, doi: 10.1080/10095020.2021.1957723.

[13] J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke, "Gradient boosting for extreme quantile regression," *Extremes*, vol. 26, no. 4, pp. 639–667, Jan. 2023, doi: 10.1007/s10687-023-00473-x.

[14] E. Bueechi *et al.*, "Crop yield anomaly forecasting in the Pannonian basin using gradient boosting and its performance in years of severe drought," *Agric. For. Meteorol.*, vol. 340, p. 109596, Jan. 2023, doi: 10.1016/j.agrformet.2023.109596.

[15] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci. Rep.*, vol. 12, no. 1, p. 6256, Jan. 2022, doi: 10.1038/s41598-022-10358-x.

[16] D. M. Atallah, M. Badawy, A. El-Sayed, and M. A. Ghoneim, "Predicting kidney transplantation outcome based on hybrid feature selection and KNN classifier," *Multimed. Tools Appl.*, vol. 78, no. 14, pp. 20383–20407, Jan. 2019, doi: 10.1007/s11042-019-7370-5.

[17] P. W. Khan, S.-J. Park, S.-J. Lee, and Y.-C. Byun, "Electric Kickboard Demand Prediction in Spatiotemporal Dimension Using Clustering-Aided Bagging Regressor," *J. Adv. Transp.*, vol. 2022, pp. 1–15, Jan. 2022, doi: 10.1155/2022/8062932.

[18] S. B. Xu, S. Y. Huang, Z. G. Yuan, X. H. Deng, and K. Jiang, "Prediction of the Dst Index with Bagging Ensemble-learning Algorithm," *Astrophys. J. Suppl. Ser.*, vol. 248, no. 1, p. 14, Jan. 2020, doi: 10.3847/1538-4365/ab880e.

[19] H. Suryono, H. Kuswanto, and N. Iriawan, "Two-Phase Stratified Random Forest for Paddy Growth Phase Classification: A Case of Imbalanced Data," *Sustainability*, vol. 14, no. 22, p. 15252, Nov. 2022, doi: 10.3390/su142215252.

[20] M. Aljabri *et al.*, "Machine Learning-Based Detection for Unauthorized Access to IoT Devices," *J. Sens. Actuator Networks*, vol. 12, no. 2, p. 27, Jan. 2023, doi: 10.3390/jsan12020027.

[21] S. Dhaliwal, A.-A. Nahid, and R. Abbas, "Effective Intrusion Detection System Using XGBoost," *Information*, vol. 9, no. 7, p. 149, Nov. 2018, doi: 10.3390/info9070149.

[22] A. Nagaraju, M. A. Kumar Reddy, C. Venugopal Reddy, and R. Mohandas, "Multifactor Analysis to Predict Best Crop using Xg-Boost Algorithm," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, Nov. 2021, pp. 155–163. doi: 10.1109/ICOEI51242.2021.9452918.

[23] S. Hazra, S. Karforma, A. Bandyopadhyay, S. Chakraborty, and D. Chakraborty, "Prediction of Crop Yield Using Machine Learning Approaches for Agricultural Data," Nov. 2023. doi: 10.36227/techrxiv.23694867.v1.

[24] J. Ge *et al.*, "Prediction of Greenhouse Tomato Crop Evapotranspiration Using XGBoost Machine Learning Model," *Plants*, vol. 11, no. 15, p. 1923, Nov. 2022, doi: 10.3390/plants11151923.

[25] A. Gupta and A. Singh, "Prediction Framework on Early Urine Infection in IoT–Fog Environment Using XGBoost Ensemble Model," *Wirel. Pers. Commun.*, vol. 131, no. 2, pp. 1013–1031, Nov. 2023, doi: 10.1007/s11277-023-10466-5.

[26] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Sci. Rep.*, vol. 13, no. 1, p. 22588, Jan. 2023, doi: 10.1038/s41598-023-49962-w.

[27] D. Zeppilli, G. Ribaudo, N. Pompermaier, A. Madabeni, M. Bortoli, and L. Orian, "Radical Scavenging Potential of Ginkgolides and Bilobalide: Insight from Molecular Modeling," *Antioxidants*, vol. 12, no. 2, p. 525, Jan. 2023, doi: 10.3390/antiox12020525.

[28] L. S. Cedric *et al.*, "Crops yield prediction based on machine learning models: Case of West African countries," *Smart Agric. Technol.*, vol. 2, p. 100049, Jan. 2022, doi: 10.1016/j.atech.2022.100049.

[29] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.