ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1576

Predicting Arabica Production: Data Mining Approach Using Linear Regression and KNN Algorithms

Sheilu Amor Q. Wawa[1]

[1]Sultan Kudarat State University Isulan, Sultan Kudarat

[1]https://orcid.org/0009-0003-0623-0122

Publication Date: 2026/08/27

Abstract: Arabica coffee is one of the most economically significant agricultural commodities worldwide, with production influenced by environmental, agronomic, and socio-economic factors. Accurate prediction of coffee yield is essential for improving productivity and sustainability in the agricultural sector. This study adopts a data mining approach to predict Arabica coffee production using two machine learning algorithms: Linear Regression and K-Nearest Neighbor (KNN). The study utilizes secondary agricultural datasets and applies data preprocessing techniques, including cleaning, normalization, and feature selection. Predictive models were developed and evaluated using standard performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Findings reveal that both algorithms can effectively model coffee production patterns, with variations in performance depending on data distribution and complexity. The results demonstrate the potential of data mining techniques in enhancing agricultural forecasting and supporting data-driven decision-making for coffee production management.

Keywords: *Arabica Coffee Production, Machine Learning, Linear Regression, K-Nearest Neighbor, Predictive Analytics.*

How to Cite: Sheilu Amor Q. Wawa (2026) Predicting Arabica Production: Data Mining Approach Using Linear Regression and KNN Algorithms. *International Journal of Innovative Science and Research Technology*, 11(5), 4655-4661. https://doi.org/10.38124/ijisrt/26may1576

I. INTRODUCTION

Arabica coffee is one of the most widely consumed beverages in the world, with a global market valued in billions of dollars annually. As a major agricultural commodity, it plays a crucial role in the economies of many developing countries, including the Philippines. The increasing global demand for high-quality coffee has intensified research efforts aimed at understanding the factors that influence its production, quality, and overall yield [4]. However, coffee production systems are increasingly affected by climate variability, which poses significant challenges to yield stability and long-term sustainability [1].

The production of Arabica coffee is influenced by a wide range of environmental and agronomic factors, including climate conditions, soil characteristics, and farming practices. Variables such as temperature, rainfall, soil fertility, and crop management techniques significantly affect both the quantity and quality of coffee yield [3], [7]. These factors interact in complex and often non-linear ways, making coffee production a dynamic and challenging agricultural process. In addition, global supply chain dynamics and market competitiveness further complicate production planning and resource allocation [6], [8]. These complexities highlight the need for advanced analytical approaches capable of extracting meaningful insights from agricultural data.

In recent years, the application of machine learning and data mining techniques in agriculture has gained significant attention due to their ability to model complex relationships and generate accurate predictions. Data mining enables the extraction of hidden patterns and knowledge from large datasets, making it particularly suitable for agricultural forecasting and decision support systems. Previous studies have demonstrated the effectiveness of machine learning approaches in various domains, including predictive maintenance and smart manufacturing, where data-driven techniques improve efficiency and system performance [2]. Similarly, in agriculture, modern data-driven technologies have been utilized to enhance coffee production systems and optimize supply chain operations [5].

Among the various predictive techniques, Linear Regression and K-Nearest Neighbor (KNN) are widely used algorithms in data mining and machine learning. Linear Regression is a statistical method that models the relationship

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1576

between a dependent variable and one or more independent variables, enabling researchers to quantify the impact of influencing factors on coffee yield. On the other hand, K-Nearest Neighbor is a non-parametric supervised learning algorithm that predicts outcomes based on the similarity between data points. It operates by identifying the nearest neighbors of a given observation and using their values to estimate the target variable.

The integration of these algorithms into agricultural data analysis reflects the growing trend of utilizing advanced analytics to optimize production systems. By leveraging predictive models, researchers and stakeholders can gain deeper insights into the critical factors affecting coffee production, enabling the development of more effective, data-driven strategies for improving productivity and sustainability. Furthermore, predictive analytics supports better resource allocation, reduces uncertainty, and enhances decision-making processes in agricultural operations.

Despite the growing adoption of machine learning techniques in agriculture, there remains a need to evaluate and compare the effectiveness of different predictive models in specific agricultural contexts. Given the multifactorial nature of Arabica coffee production, a comprehensive analytical approach is necessary to address the challenges faced by coffee farmers. Thus, this study adopts a data mining approach to predict Arabica coffee production using Linear Regression and K-Nearest Neighbor algorithms, with the aim of assessing their predictive performance and applicability in agricultural forecasting.

➢ *Conceptual Framework*

II. METHODOLOGY

➢ *Research Design*

This study adopts a quantitative research design using a data mining approach to predict Arabica coffee production. The study focuses on comparing the predictive performance of two machine learning algorithms, namely Linear Regression (LR) and K-Nearest Neighbor (KNN), in estimating coffee yield based on selected variables.

A data-driven framework was employed, consisting of data collection, preprocessing, model development, and evaluation. This approach enables the extraction of meaningful patterns from agricultural data and supports predictive modeling for decision-making.

In this study, the open-source distribution of Python version 3.11 and Jupyter Notebook version 4.0 for data science and machine learning were utilized to predict the arabica coffee production. The hardware and software were identified, which are all necessary to materialize the study. These materials have an important role in predicting the production of coffee as mentioned in the conceptual framework.

The dataset has 1,312 columns and 14 rows that consist of company, producer, mill, region, ID, ICO, owner, species, altitude, grading date, harvest year, number of bags, farm, and country.

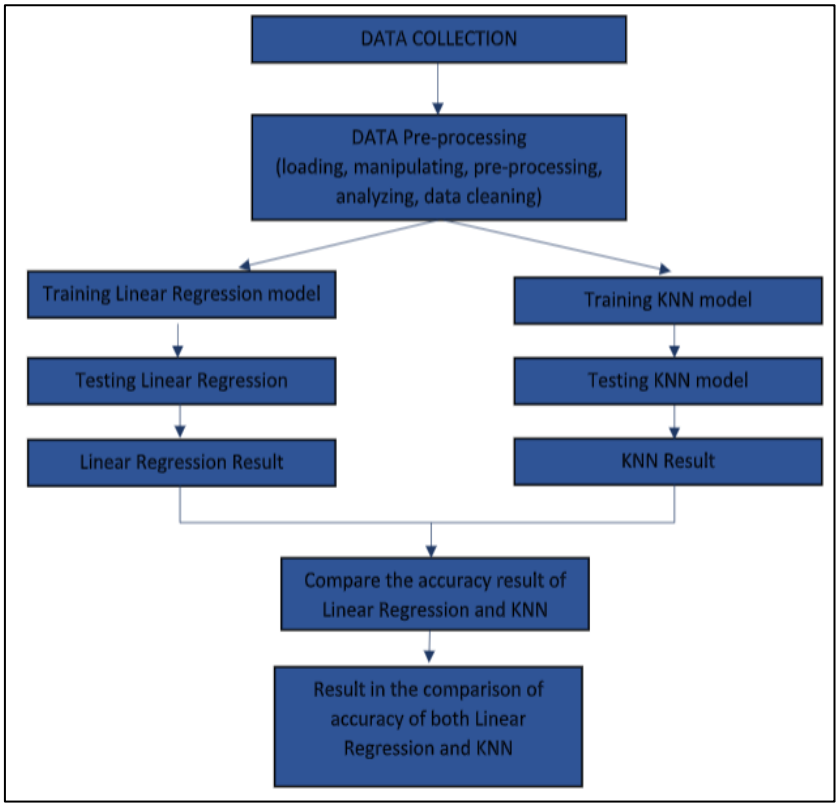

Fig 1 Conceptual Framework of the Study

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1576

The conceptual framework of the study illustrates the overall process of predicting Arabica coffee production using data mining techniques. The process begins with data collection, where the dataset is obtained from a secondary source. This is followed by data preprocessing, which includes data loading, cleaning, transformation, and feature selection to ensure data quality and relevance.

After preprocessing, the dataset is divided into training and testing sets. Two predictive models, Linear Regression and K-Nearest Neighbor are then developed and trained using the processed data. Each model is subsequently tested to generate prediction outputs. The results of both models are evaluated using regression error metrics, and their predictive performances are compared to determine which algorithm provides more accurate predictions for Arabica coffee production.

➢ *Dataset Description*

The dataset used in this study was obtained from Kaggle and contains information related to Arabica coffee production. The dataset originally consists of 1,312 observations and 14 variables, including attributes such as producer, farm, country, ICO ID, owner, mill, altitude, region, harvest year, grading date, species, and number of bags.

After preprocessing, nine (9) relevant variables were retained for analysis. The variable "number of bags" (NumBag) was used as the dependent variable representing coffee production, while the remaining variables served as independent predictors.

➢ *Data Preprocessing*

Data preprocessing was conducted to prepare the dataset for analysis and modeling. The dataset was obtained from Kaggle and consists of 1,312 observations and 14 variables. Initial processing involved examining the relationships between independent variables and the dependent variable.

Categorical variables were transformed into numerical representations through factorization to enable compatibility with machine learning algorithms. Subsequently, irrelevant and redundant features, including grading date, ICO, harvest year, altitude, and ID, were removed using feature selection techniques. The dependent variable, originally labeled as "number of bags," was renamed to NumBag for consistency and clarity. Finally, the dataset was reorganized by positioning the dependent variable at the last column to facilitate model training and evaluation.

➢ *Model Development*

Two machine learning algorithms were implemented in this study: Linear Regression and K-Nearest Neighbor.

Linear Regression was used to model the relationship between the dependent variable and multiple independent variables by fitting a linear equation to the observed data.

- *Formula 1*

  Linear Regression

$$Y = mx^1 + mx^2 + mx^3 + b$$

On the other hand, the K-Nearest Neighbor (KNN) algorithm predicts the value of a data point based on the average values of its nearest neighbors, using similarity measures between observations.

- *Formula 2*

  K-Nearest Neighbor

$$\mu\hat{}(\sim x_o) = 1/k \sum (j=1 \text{ to } k)\ yNN(\sim x_o, j)$$

These models were selected due to their ability to capture both linear and non-linear relationships within the dataset.

➢ *Model Training*

Following model development, the dataset was partitioned into training and testing subsets using the train_test_split method, with 70% of the data allocated for training and 30% reserved for testing. This separation ensures that the models are evaluated on unseen data, providing a more reliable estimate of their predictive performance.

During the training phase, the algorithms learn the relationship between the independent variables and the dependent variable (NumBag). For Linear Regression, the training process involves estimating the regression coefficients that minimize the difference between predicted and actual values. This is typically achieved by minimizing the sum of squared errors, allowing the model to identify the linear relationship that best fits the data.

In contrast, the K-Nearest Neighbor algorithm does not construct an explicit model during training. Instead, it stores the training data and relies on distance-based computations during prediction. When a new input is introduced, the algorithm identifies the k closest data points in the training set and computes the predicted value based on the average of these neighbors. As such, the effectiveness of KNN depends on the structure and distribution of the training data.

The use of separate training and testing datasets ensures that the models are not evaluated on the same data used for learning, thereby reducing the risk of overfitting and improving the generalizability of the results.

➢ *Model Evaluation*

Model performance was evaluated using standard regression error metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics were used to measure the discrepancies between predicted and actual values, providing a comprehensive assessment of model accuracy.

MSE measures the average squared difference between predicted and actual values, while RMSE represents the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1576

standard deviation of prediction errors. MAE calculates the average absolute deviation, and MAPE expresses prediction error as a percentage. Lower values across these metrics indicate better model performance and higher predictive accuracy.

➢ *Tools and Implementation*

The study was implemented using the open-source distribution of Python (version 3.11) and Jupyter Notebook version 4.0 for data science and machine learning. Libraries such as Pandas and NumPy were used for data manipulation, while Scikit-learn was utilized for model development, training, and evaluation.

III. RESULTS (FINDINGS) AND DISCUSSION

➢ *Dataset Overview*

The preprocessed dataset consists of 1,312 observations and nine selected variables used for model training and evaluation. The dependent variable, NumBag, represents Arabica coffee production. Data inspection confirmed that the dataset contains no missing values, indicating its suitability for predictive modeling.

Table 1 Presents the Summary of the Preprocessed Data with Number of Non-Null Values

Column	Non-Null Count	Data Type
Species	716	Object
Owner	716	Object
Country	716	Object
Farm	716	Object
Mill	716	Object
Company	716	Object
Region	716	Object
Producer	716	Object
NumBag	716	Integer

The dataset used in this study consists of 1,312 observations and 14 variables. Following data preprocessing, nine relevant features were retained for analysis. The variable NumBag was designated as the dependent variable, representing Arabica coffee production, while the remaining variables were treated as independent predictors.

➢ *Model Performance Evaluation*

Table 2 Metric Errors Results of Both Linear Regression and K-Nearest Neighbor

Model Name	Linear Regression	K-Nearest Neighbor
MSE	16474.63	15489.02
RMSE	128.35	124.45
MAPE	15.33%	11.63%
MAE	110.57	94.82

Table 2 shows the error results of both Linear Regression and K-Nearest neighbor with the metric errors namely MSE, RMSE, MAE and MAPE. The MSE measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values.

In Table 2 the K-Nearest Neighbor model got the lowest mean squared error total of 15489.02, the Linear Regression got 16474.63.

The RMSE measures the average difference between values predicted by a model and the actual values. It is the standard deviation of the residuals, which indicates average model prediction error. The lower values indicate a better fit. Shweta Gupta (2021). Table 2 shows that the RMSE of Linear Regression is 128.35 followed by a K-Nearest Neighbor total of 124.5. Since the KNN has a lowest RMSE score than linear regression, the RMSE was 124.5. So far, KNN seems better than linear regression for making predictions on this data (Anshul Kumar 2021).

Moreover, moving to MAPE is an accuracy measure based on the relative percentage of errors considering that the K-nearest neighbor obtained 11.63% and linear regression got 15.33%. The KNN always performs better than the Linear regression model, therefore it is reasonable to choose the knn model as the regression model to predict the mape of the datasets (li kuang, han yan, yujia zhu, shenmei tu & xiaoliang fan 2019). Lastly, the MAE also known as L1 is a row-level error calculation where the non-negative difference between the prediction and the actual is calculated. MAE is the aggregated mean of these errors, which helps us understand the model performance over the whole dataset as given in table 2 k-nearest neighbor obtained a score of 94.82 while linear regression got 110.57. Among the two models, the k-nearest neighbor is closer to MAE. It is reasonable that the KNN was more accurate the model since it is closer to zero.

➢ *Comparative Analysis of Models*

The superior performance of KNN suggests that the dataset contains non-linear relationships between the predictor variables and coffee production. Linear Regression assumes a linear relationship, which limits its ability to capture complex patterns in the data. In contrast, KNN relies on similarity-based prediction, allowing it to adapt to non-linear structures.

Furthermore, the lower error values of KNN indicate that it better captures variations in production across different regions and conditions. This demonstrates that non-parametric models are more suitable for datasets with heterogeneous characteristics.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1576

➢ *Residual Analysis*

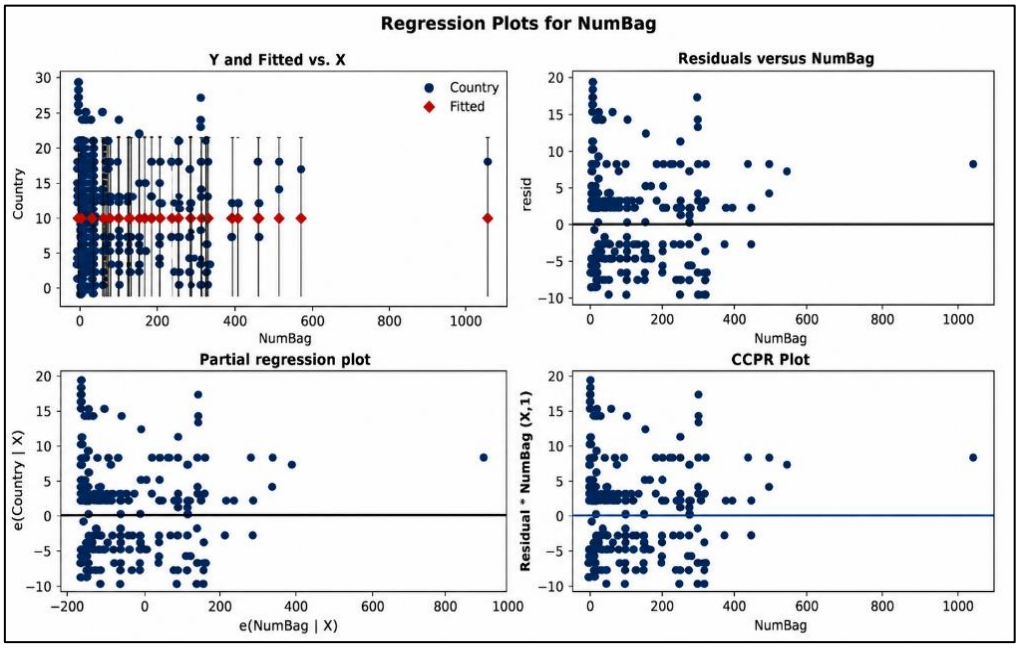

Fig 2 Regression Plot and Residual Distribution for Linear Regression

Additionally, the model score of Linear Regression is only 10% while the KNN for the training got 45% and for the testing is only 6%. After plotting the y test and residuals of a linear regression model, it is observed that the scatter plot does not show a random distribution of points around the horizontal line, it could indicate that the model is not the best fit for the dataset. If the residuals show a pattern, such as a curved or funnel-shaped pattern, it may suggest that the model is not capturing all the relevant information in the data, leading to a poor fit.

➢ *Prediction Comparison*

Table 3 Comparison of Actual and Predicted Values

Linear Regression	**Actual Value**	**Predicted Value**	**Difference**
0	250	206.340899	43.659101
1	125	212.764679	-87.764679
2	320	210.599712	109.400288
3	250	106.676539	143.323461
4	100	208.969628	108.969628
5	1	206.168677	-205.168677
6	29	182.940118	-153.940118
7	1	171.216998	-170.216998
8	10	55.047199	-45.047199
9	320	209.717674	110.282326
K-Nearest Neighbor	**Actual Value**	**Predicted Value**	**Difference**
0	250	250.0	0.0
1	125	204.2	-79.2
2	320	320.0	0.0
3	250	197.0	53.0
4	100	267.0	-167.0
5	1	191.4	-190.4
6	29	130.6	-101.6
7	29	160.8	-159.8
8	10	57.0	-47.0
9	320.0	320.0	0.0

The table 3 reveals the residuals of the actual value and predicted value. The standard error of the regression, frequently referred to standard error to estimate, and measuring the average distance between the values that were

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1576

observed and the regression line, to lead the outcome in different actual value. It utilize the response of variable units conveniently to show how inaccurate the regression model is generally.

➢ *Discussion*

The results of this study demonstrate that the K-Nearest Neighbor (KNN) algorithm consistently outperforms Linear Regression in predicting Arabica coffee production. This is evidenced by lower error values across all evaluation metrics, including MSE, RMSE, MAE, and MAPE. Lower values in these metrics indicate that the predicted outputs are closer to the actual values, confirming the superior predictive performance of KNN.

The observed performance difference can be attributed to the nature of the dataset. Arabica coffee production is influenced by multiple interacting factors such as region, farm characteristics, and production conditions, which may exhibit non-linear relationships. Linear Regression assumes a linear relationship between the dependent and independent variables, limiting its ability to model such complexities. In contrast, KNN is a non-parametric algorithm that does not impose strict assumptions on the data, allowing it to capture non-linear patterns through similarity-based prediction.

Residual analysis further supports these findings. The Linear Regression model exhibits larger deviations between predicted and actual values, with residual patterns indicating that the model does not adequately fit the data. This suggests that important relationships within the dataset are not captured by the linear model. On the other hand, the KNN model produces smaller residuals, indicating improved accuracy and consistency in predictions.

These findings are consistent with prior studies that highlight the effectiveness of non-parametric and instance-based learning algorithms in handling complex and heterogeneous datasets, particularly in agricultural applications. The flexibility of KNN in adapting to local data patterns makes it more suitable for modeling real-world agricultural data, which often involve variability and non-linear interactions.

However, despite its superior accuracy, KNN has certain limitations. Its performance is highly dependent on the selection of the parameter k and the distribution of the dataset. Additionally, KNN can be computationally intensive when applied to large datasets, as it requires storing and processing all training data during prediction. In contrast, Linear Regression offers advantages in terms of simplicity, interpretability, and computational efficiency.

Overall, the results emphasize the importance of selecting appropriate predictive models based on data characteristics. While Linear Regression provides a straightforward approach for modeling linear relationships, KNN offers a more flexible and accurate alternative for datasets with complex structures, such as Arabica coffee production data.

IV. CONCLUSION AND RECOMMENDATIONS

This study applied a data mining approach to predict Arabica coffee production using Linear Regression and K-Nearest Neighbor (KNN) algorithms. The dataset, consisting of 1,312 observations and multiple production-related variables, was preprocessed and used to develop and evaluate predictive models.

The results indicate that the KNN model outperforms Linear Regression across all evaluation metrics, including MSE, RMSE, MAE, and MAPE. The superior performance of KNN suggests that the dataset contains non-linear relationships that are not effectively captured by Linear Regression. As a non-parametric algorithm, KNN demonstrates greater flexibility in modeling complex patterns within the data, leading to improved prediction accuracy.

The findings of this study highlight the effectiveness of data mining techniques in agricultural forecasting. By applying appropriate machine learning models, it is possible to generate more accurate predictions of coffee production, which can support better decision-making for farmers, stakeholders, and agricultural planners.

Despite these contributions, the study is limited by the use of a secondary dataset and the application of only two predictive models. Future research may explore the use of additional machine learning algorithms, larger and more diverse datasets, and the inclusion of external factors such as climate variables to further improve prediction performance.

In conclusion, the study demonstrates that KNN is a more suitable model for predicting Arabica coffee production under the given dataset conditions, providing a valuable framework for future research and practical applications in agricultural analytics.

STATEMENTS AND DECLARATIONS

➢ *Funding Details:*

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

➢ *Disclosure Statement:*

The authors declare that there are no competing financial or non-financial interests that could have appeared to influence the work reported in this paper.

➢ *Declaration of Generative AI in Scientific Writing:*

During the preparation of this work, the author(s) used ChatGPT by OpenAI to assist in improving language clarity, grammar, organization, and academic writing structure. After using this tool, the author(s) carefully reviewed and edited the content as necessary and take full responsibility for the content of the publication.

➢ *Acknowledgement:*

The authors would like to express their sincere gratitude to Sultan Kudarat State University – Isulan Campus and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1576

University of Immaculate Conception for the academic support and guidance provided throughout the conduct of this study. Appreciation is also extended to all individuals and organizations who contributed valuable insights and assistance to the completion of this research.

➢ *Ethical Approval:*

This study utilized secondary data obtained from publicly accessible datasets and did not involve human participants, animals, or confidential personal information. Therefore, formal ethical approval was not required for this research.

REFERENCES

➢ *Journal Article with DOI*

[1]. Bunn, C., Läderach, P., Rivera, O., & Kirschke, D. (2015). A bitter cup: Climate change profile of global production of Arabica and Robusta coffee. Climatic Change, 129(1–2), 89–101. https://doi.org/10.1007/s10584-014-1306-x

[2]. Duarte, A., Uribe, J. C., Sarache, W., & Calderón, A. J. E. (2019). Economic, environmental, and social assessment of bioethanol production using multiple coffee crop residues. Energies, 12(2), 332. https://doi.org/10.3390/en12020332

[3]. Effendi, D., & Rismaya, M. (2020). Design and development of coffee production information system to support coffee production productivity in farmers group. IOP Conference Series: Materials Science and Engineering, 725(1), 012052. https://doi.org/10.1088/1757-899X/725/1/012052

[4]. Kittichotsatsawat, Y., Jangkrajrang, V., & Tippayawong, K. Y. (2020). Enhancing coffee supply chain towards sustainable growth with big data and modern agricultural technologies. Sustainability, 12(4), 1634. https://doi.org/10.3390/su12041634

[5]. Torok, A., Mizik, T., & Jambor, A. (2019). The competitiveness of global coffee trade. Journal of International Studies, 12(2), 127–141. https://doi.org/10.14254/2071-8330.2019/12-2/8

➢ *Journal Article Without DOI (When DOI is not Available)*

[6]. Çinar, Z. M., et al. (2020). Machine learning in predictive maintenance toward sustainable smart manufacturing in the industry. International Journal of Engineering and Issues, 4(6), 150–156.

[7]. Saragih, J. R. (2013). Socioeconomic and ecological dimension of certified and conventional Arabica coffee production in North Sumatra, Indonesia. Asian Journal of Agriculture and Rural Development, 3(3), 93–107.

➢ *Online Discussion Paper*

[8]. Lewin, B., Giovannucci, D., & Varangis, P. (2004). Coffee markets: New paradigms in global supply and demand. World Bank Agriculture and Rural Development Discussion Paper.

➢ *Online Website*

[9]. International Coffee Organization. (n.d.). Total production by all exporting countries. Retrieved from https://www.ico.org/prices/po-production.pdf

BIONOTE

Sheilu Amor Wawa is a faculty member and Campus Production and Income Generating Projects (IGP) Coordinator at Sultan Kudarat State University – Isulan Campus. She earned her Bachelor of Science in Accountancy and Master of Business Administration degrees and is currently pursuing a Doctor in Business Management major in Information Systems at the University of the Immaculate Conception. Her work focuses on accounting education, business analytics, information systems, technopreneurship, and institutional digital transformation research.