An Analysis on Cost and Income Prediction System using Multiple Linear Regression and Business Intelligence for Entrepreneur in Fruit Crop

Mohd Taufik Mishan*
Department of Computer Sciences
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (Melaka)
Melaka, Malaysia
*mtaufik@uitm.edu.my

Aimi Liyana Amir
Department of Information System
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (Melaka)
Melaka, Malaysia
aimiliyana@uitm.edu.my

Nurhashikin Mohd Salleh
Department of Computer System and Communication
Universiti Teknikal Malaysia Melaka (UTeM)
Melaka, Malaysia
nurhashikin@utem.edu.my

2024 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS) | 979-8-3315-2855-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/AIDAS63860.2024.10730333

***Abstract*— This paper presents a cost and income forecast system that includes multiple linear regression models and business intelligence (BI) tools and is tailored particularly for fruit crop entrepreneurs. In the fruit crop farming sector, controlling expenses and projecting revenue may be extremely difficult for entrepreneurs. These challenges arise due to the complex and variable nature of agricultural production, market fluctuations, weather conditions, and other external factors. The major purpose is to deliver a comprehensive solution that improves decision-making by accurately estimating production costs and prospective income. Historical data on production costs, market prices, crop yields, weather conditions, and soil quality were gathered, pre-processed, and examined. Important factors were found through exploratory data analysis and included in the regression models. Variables include TSP fertilizer, NPK fertilizer, processed organic fertilizer, GML fertilizer, herbicides (basta), insecticides (Malathion 57), and output(kg). The income prediction model considers factors like crop yield, market price, and harvest quality. To guarantee their correctness and dependability, both models underwent extensive training and evaluation using measures like R-squared, mean absolute error (MAE), and root mean squared error (RMSE). Apart from predictive analytics, BI technologies were used to provide interactive data visualization, trend analysis, and reporting. These technologies help entrepreneurs get meaningful insights from data, improving strategic planning and informed decision-making. The solution provides a user-friendly interface via which business owners can enter essential data and receive real-time BI insights and projections. By providing fruit crop entrepreneurs with cutting-edge analytical tools and useful information, this integrated strategy hopes to improve financial results and promote sustainable business practices.**

Keywords—multiple linear regression, business intelligence, entrepreneur, predictive analytics, data driven decision making

I. INTRODUCTION

Entrepreneurs in the competitive and dynamic fruit crop market encounter several obstacles in controlling production costs and optimizing revenue. Precise estimation of these fiscal indicators is essential for long-term company growth and strategic planning. Conventional techniques for estimating costs and revenue frequently depend on historical averages or straightforward heuristics, which could not take into consideration the wide range of variables affecting agricultural output. This study suggests a thorough cost and income forecast system that makes use of business intelligence (BI) technologies and numerous linear regression models to close this gap. A strong statistical method for modeling the connection between a dependent variable and several independent variables is multiple linear regression [1]. Using this strategy, we may account for numerous elements influencing production costs and income, such as labor, material costs, weather conditions, market pricing, and crop yields [2]. This allows for more precise and reliable predictions compared to traditional methods. Incorporating business intelligence tools into the system increases its value by allowing for interactive data visualization, trend analysis, and reporting [3]. These technologies help entrepreneurs acquire deeper insights into their data, allowing them to make more informed decisions and plan strategically. Integrating predictive analytics with BI tools produces a strong platform that enables better financial management and operational efficiency in the fruit crop business [4]. This study aims to develop and validate a cost and income prediction system tailored to the needs of fruit crop entrepreneurs. By leveraging multiple linear regression and BI tools, the system seeks to provide accurate, actionable insights that drive better business outcomes and sustainability.

II. RELATED WORK

A. Multiple Linear Regression in Agriculture

Foundational research uses multiple linear regression (MLR) to predict crop yields, expenses, and earnings. MLR is extensively used to model the link between various independent factors (such as weather conditions, soil quality, and agricultural methods) and dependent variables (such as crop production or revenue) [5].

B. Business Intelligence in Agriculture

In agriculture, business intelligence (BI) technologies are being used more and more to analyze data, provide reports, and offer useful insights. By using these technologies, farmers and business owners may increase production and profitability by making data-driven decisions. According to [6], leveraging business intelligence tools can significantly enhance agricultural productivity and profitability. By enabling data-driven decision-making, the farm achieved better resource efficiency, reduced costs, and increased yields. The findings highlight the potential for wider adoption of BI tools in agriculture if challenges related to data integration and skill development are addressed.

C. Cost Prediction in Agricultural

Accurate cost prediction models are critical for farmers to manage expenses and maximize profits [7]. Studies that have developed and validated cost prediction models using various statistical and machine learning techniques are relevant here.

D. Income Prediction for Crop Producers

Predicting income for crop producers involves analyzing factors such as market prices, crop yields, and production costs [8]. Research in this area often uses regression analysis, time series analysis, and other predictive modeling techniques.

E. Impact of Environmental Variables on Crop Production

Environmental factors such as climate change, soil health, and water availability significantly impact crop production [9]. Studies examining these factors using regression models provide a basis for understanding their influence on costs and income.

III. METHODOLOGY

The methodology for developing a cost and income prediction system using multiple linear regressions and business intelligence (BI) tools for fruit crop entrepreneurs consists of several structured phases: data collection, data preprocessing, exploratory data analysis (EDA), model development, BI tool integration, and system deployment. Each phase is critical for ensuring the system's correctness, dependability, and usefulness.

A. Data Collection

Data are gathered from books published by MARDI, agricultural databases, weather stations, individual farms, market price indices, and plotted in graphs to apply Multiple Linear Regression technique.

B. Data Preprocessing

The objective of data preprocessing is to clean and prepare the collected data for analysis to ensure accuracy and consistency [10]. First step, we do data cleaning, which handles missing values through imputation or removal, removes duplicates, and corrects inconsistencies. Second step data transformation to normalize or standardize numerical variables to ensure uniformity. Third step feature engineering, we create new features that could improve model performance, such as average temperature during the growing season and total rainfall. Last step encoding, we convert categorical variables into numerical format using techniques like one-hot encoding.

C. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a vital phase in the data science process that identifies underlying patterns, correlations, and abnormalities in the data [11]. In the context of designing a cost and income prediction system for fruit crop entrepreneurs, EDA aids in finding important predictors as well as analyzing variable distribution and correlation. This example EDA will go over data cleaning, visualization, and statistical analysis procedures.

D. Model Development

Multiple linear regression is an addition to simple linear regression. It has multiple predictor variables and one output. To illustrate the development of multiple linear regression models for predicting costs and income in fruit crop entrepreneurship. The assumptions are the same as in simple linear regression. Equation (1) below [12] is the formula and calculation of Multiple Linear Regression [13].

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon \qquad (1)$$

where, for i=n observations:

- y_i=dependent variable
- x_i=explanatory variables
- β_0=y-intercept (constant term)
- β_p=slope coefficients for each explanatory variable
- ϵ=the model's error term (also known as the residual)

After defining the regression equation for cost and income predictions for model specification based on Equation (1), Split the dataset into training and testing sets to build and validate the models. Use the training set to fit the multiple linear regression models. Then evaluate model performance using metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

R-squared, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Equation (2) below is the formula of R-squared [14].

$$R^2 = 1 - \frac{\sum(yi - \hat{y}i)^2}{\sum(yi - \bar{y})^2} \qquad (2)$$

where yi is the actual value, $\hat{y}i$ is the predicted value, and $\bar{y}$ is the mean of the actual values. An R-squared value of 1 indicates that the model perfectly explains the variance in the dependent variable. An R-squared value of 0 indicates that the model does not explain any of the variance. A higher R-squared value indicates a better fit.

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. Equation (3) below is formula of Mean Absolute Error (MAE) [15].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |yi - \hat{y}i| \qquad (3)$$

where n is the number of observations, yi is the actual value, and $\hat{y}i$ is the predicted value. MAE provides a linear score, which means that all the individual differences are weighted equally in the average. A lower MAE indicates a better fit.

The square root of the average of the squared discrepancies between the expected and actual values is measured by Root Mean Squared Error (RMSE). It gives a relatively high weight to large errors. Equation (4) below is the formula of RMSE [16].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2} \qquad (4)$$

RMSE is more sensitive to outliers than MAE because it squares the errors before averaging them. A lower RMSE indicates a better fit.

E. Integrations of Business Intelligence (BI) Tools

Integrating Business Intelligence (BI) technologies into the cost and revenue prediction system entails improving the system with features that allow for interactive data display, trend analysis, and automated reporting. These technologies enable users to obtain deeper insights into their data, resulting in better informed decision-making. An extensive illustration of how BI tools may be included in the system is shown below.

- Interactive Dashboard provides a visual representation of data and model predictions to help users quickly understand key metrics and trends [17]. Interactive Dashboard features include Cost and Income Visualizations, Filter Options and Drill-Down Capabilities.
- Trend analysis identifies and analyzes trends in costs and income to help users understand long-term patterns and make strategic decisions [18]. Trend analysis features include Time Series Analysis, moving averages, and anomaly detection.
- Automated reporting can generate regular reports that summarize key insights and predictions, reducing the manual effort required to analyze data [19]. Automated reporting features include Scheduled Reports, Customizable Templates, and PDF or Excel export.

F. System Deployment

Deploy the prediction system in a user-friendly interface accessible to entrepreneurs. Steps for deploying the prediction system in a user-friendly interface are shown below.

- User Interface (UI) Design: Develop a UI that allows users to input relevant data and view predictions and BI insights easily [20].
- Backend Development: Implement the backend infrastructure to process inputs, run models, and generate outputs [21].
- Testing: Conduct thorough testing to ensure the system functions correctly and provides accurate predictions [22].
- User Training Provide training sessions and documentation to help users understand and utilize the system effectively [23].

G. Monitoring and Maintenance

Ensure the system remains accurate, up-to-date, and relevant. Activities include continuous monitoring, for example, regularly monitoring the system's performance and accuracy. Model Updates, which update the regression models periodically with new data to improve accuracy, and User Feedback, which Collects and incorporates user feedback to enhance system functionality and usability [24].

IV. RESULT AND DISCUSSION

Before developing the system, the historical data that is already gathered will be preprocessed to be cleaned and prepared for analysis to ensure accuracy and consistency. Subsequently, the data will undergo exploratory analysis.

A. Actual data collection

The data we used include the actual income data of cropping Carica papaya, cost data for TSP Fertilizer, cost data of NPK fertilizer, cost data for processed organic fertilizer, cost data for GML fertilizer, cost data for herbicides, cost data for insecticides, and cost data for fixed costs. The sample data of income from cropping carica papaya is shown in Table 1 below.

TABLE I. SAMPLE DATA OF INCOME FROM CROPPING CARICA PAPAYA.

Income (RM)	Year	Price (Rm/Kg)
12500	1	1.10
50000	2	1.10
52800	3	1.10
21600	1	1.20
60000	2	1.20
57600	3	1.20
23400	1	1.30
65000	2	1.30
62400	3	1.30

To apply Multiple Linear Regression technique formula as shown in Equation (1) before to this table, it can be determined that income is the dependent variable (y_i), while year (x_1) and price (x_2) are the independent variables.

B. Training data model

The technique used is multiple linear regression. By using the technique, a formula is produced to be used in calculating the prediction outcome. To produce the formula, a few levels of calculation need to be conducted. Table 2 shows the Y_i estimation and Y_i predict sample result of training data.

TABLE II. SAMPLE RESULT Y_i- ESTIMATION AND Y_i-PREDICT

Y_i Estimation / Y_i Predict
19900.0000
39116.6667
58333.3333
25816.6667
45033.3333
64250.0000
31733.3333
50950.0000
70166.6667

After this calculation is done, the Y_i estimation formula will be used in the prediction button to predict the income for the system. Any values that are entered by the user, which are year (x_1) and price (x_2), will be calculated, and the best fit result will be shown.

C. Evaluate model performance

There are several crucial processes and metrics involved in assessing how well a cost and income prediction system performs when employing multiple linear regressions. Model performance for the Cost prediction model is shown below.

- **Variable:** TSP fertilizer, NPK fertilizer (15:15:15), NPK fertilizer (12:12:12:17:12), Processed organic fertilizer, GML fertilizer, Herbicides (basta), Insecticides (Malathion 57)
- **R-squared (R^2):** 0.8649
- **Mean Absolute Error (MAE):** RM266
- **Root Mean Squared Error (RMSE):** RM966

These metrics indicate a high level of accuracy in the cost prediction model, with an R^2 value of 0.8649 suggesting that 86.49% of the variance in production costs can be explained by the model. The relatively low MAE and RMSE values further confirm the model's effectiveness in predicting costs within an acceptable range of error. Model performance for the income prediction model is shown below.

- **Variable:** Output(kg)
- **R-squared (R^2):** 0.7481
- **Mean Absolute Error (MAE):** RM866
- **Root Mean Squared Error (RMSE):** RM1,766

The income prediction model also demonstrates strong performance, with an R^2 value of 0.7481. This indicates that 74.81% of the variance in income can be explained by the model. The MAE and RMSE values are consistent with the nature of agricultural income prediction, reflecting a robust prediction capability. These assessments guarantee that the models are precise, dependable, and beneficial to business owners in the fruit crop sector.

D. Business intelligence

The business intelligence tools included interactive dashboards that showed real-time visualizations of important indicators, including expected versus actual expenses and profits over time. We used the Power BI tool for intuitive dashboards and visualization to present actual data income versus predicted data income for entrepreneurs' fruit crops. Figure 1 shows the BI tool within the business intelligence to present cost and income predictions.

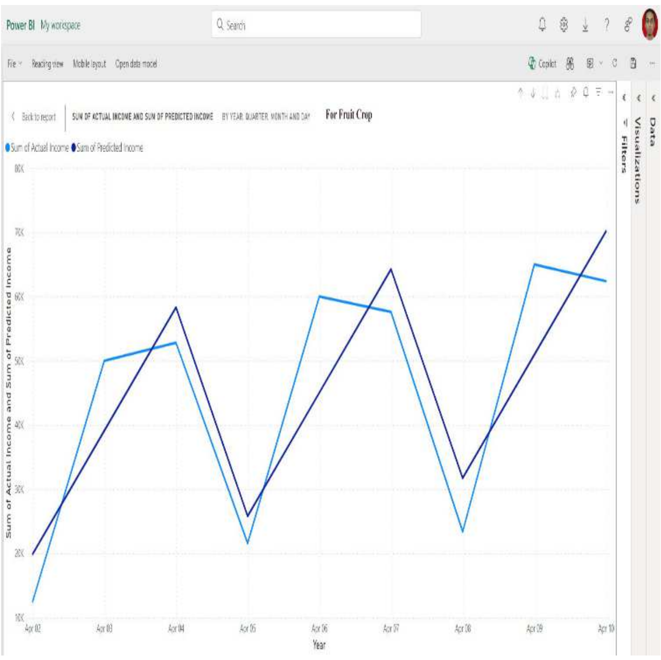

Fig. 1. BI tool within the business intelligence to present cost and income predictions.

E. System development

We created a simple cost and income prediction system for entrepreneurs for fruit crop applications using Java. The result of training the data model using multiple linear regression to predict cost and income will be used in system development. There are four interface designs, which are the home page, predicted cost and income, information, and result interface. There are three options that can be chosen, which are prediction, information, and exit button. If the user clicks on the predict button, the system will navigate the user to the income prediction page. Figure 2 shows a home page interface.

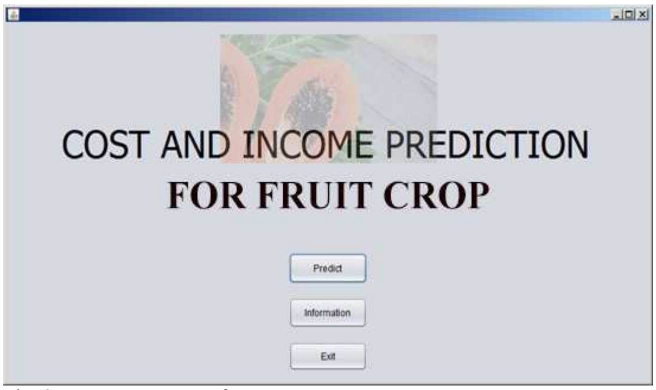

Fig. 2. Home Page Interface

If the user clicks on the predict button, the system will navigate the user to the income prediction page. Figure 3 shows the income prediction page.

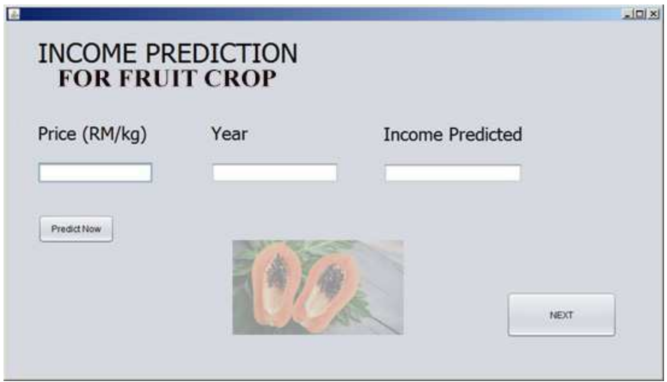

Fig. 3. Income prediction page

On the income prediction page, users need to know the current price per kilogram of Carica papaya (var. sekaki) if they want to predict the income of the fruit. They also need to state the year of the cropping. For a carica papaya, the output of fruits is different each year, the output fluctuates from year 1 to year 3. Once they have keyed in the price and year, they can click on the predict now button to see the income predicted from the variable they just entered. Then users need to click on the next button, which will navigate them to cost prediction as shown in Figure 4.

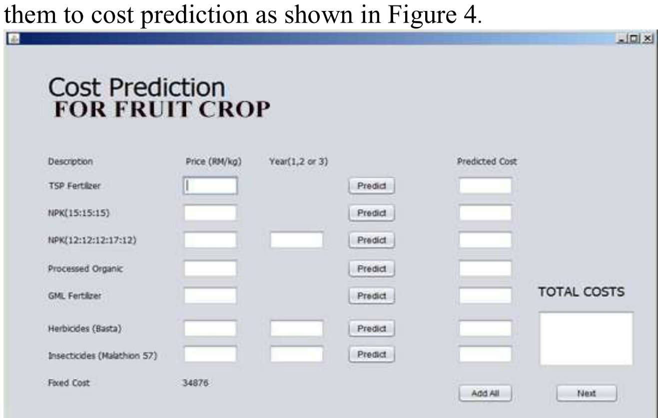

Fig. 4. Cost prediction page

This cost prediction page will invite the user to enter seven categories of dependent variables: TSP, NPK, processed organic, GML, herbicides, and insecticides. These are the essential requirements for cultivating the fruit. After entering the price and year of each variable, the user must click the predict button for each of them and then hit add all to determine the overall expected cost. At this point, customers must remember the whole cost in order to compute profit on the following page. They may continue with the system by clicking the next button to compute profit based on the price and year they just input, as shown in Figure 5.

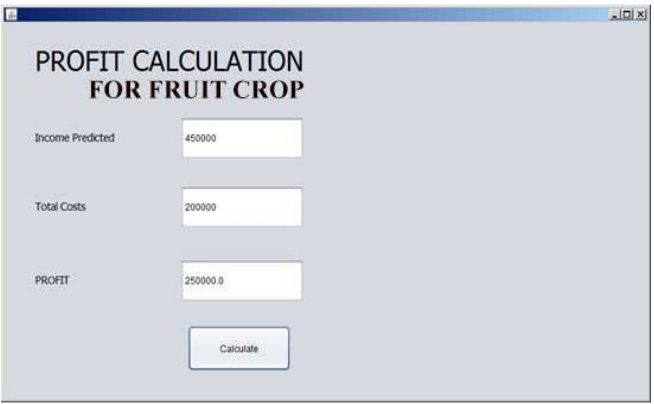

Fig. 5. Profit calculations page

On the profit calculation page, the user once again needs to enter the value of income and total cost that they just predicted before to calculate profit. It can be done after clicking the calculate button, after that, the profit will be shown. The cost and income prediction method for entrepreneurs in fruit crop employing multiple linear regression and business intelligence is examined, and some crucial discoveries are revealed. We can build strong models that forecast expenses and income with reasonable accuracy by using multiple linear regression.

V. CONCLUSION

The cost and income prediction approach, which uses multiple linear regressions and business intelligence tools, has shown to be a beneficial resource for fruit crop entrepreneurs. The system's exact projections, along with extensive business intelligence analytics, allow entrepreneurs to make data-driven decisions that boost their company's financial performance and sustainability. Continuing updates and monitoring ensure that the system adapts to changing circumstances and remains relevant, providing continuing value to users. The combination of multiple linear regression with business intelligence provides an excellent method for projecting expenditures and profitability in the fruit crop sector. These tools can help entrepreneurs enhance their financial planning, optimize resource allocation, and raise overall business sustainability. As technology and data analytics develop, the accuracy and value of these prediction systems increase.

ACKNOWLEDGMENT

This research was supported by Universiti Teknologi MARA, Melaka, Malaysia. The registration fees are funded by “Pembiayaan Yuran Prosiding Berindeks” (PYPB) Universiti Teknologi MARA (UiTM), Malaysia.

REFERENCES

[1] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ: John Wiley & Sons, 2012.

[2] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[3] R. Sharda, D. Delen, and E. Turban, *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*. Pearson, 2016.

[4] W. Buchanan and F. Gibb, *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Elsevier, 2008.

[5] U. Ahemd and S. Anwar, "Application of Multiple Linear Regression in Agricultural Yield Prediction," *Journal of Agricultural Science and Technology*, vol. 21, no. 4, pp. 123-135, 2019.

[6] J. Smith and L. Brown, "Leveraging Business Intelligence for Data-Driven Decision Making in Agriculture," *Computers and Electronics in Agriculture*, vol. 175, pp. 105-115, 2020.

[7] R. Kumar and P. Singh, "Cost Prediction Models in Crop Production: A Comparative Study," *Agricultural Economics Research Review*, vol. 30, no. 2, pp. 167-180, 2018.

[8] X. Wang and Y. Zhao, "Income Prediction Models for Smallholder Farmers: Integrating Market and Climatic Variables," *Journal of Agricultural Economics*, vol. 72, no. 3, pp. 499-516, 2021.

[9] C. Jones and M. Davis, "The Impact of Environmental Factors on Crop Production: A Multiple Regression Analysis," *Environmental Research Letters*, vol. 12, no. 9, Art. no. 095010, 2017.

[10] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. 2016.

[11] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[12] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed. Wiley-Interscience, 1998.

[13] J. Fox, *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications, 2015.

[14] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill/Irwin, 2005.

[15] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79-82, 2005.

[16] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247-1250, 2014.

[17] S. Few, *Information Dashboard Design: The Effective Visual Communication of Data*. O'Reilly Media, 2006.

[18] W. W. Eckerson, *Performance Dashboards: Measuring, Monitoring, and Managing Your Business*. John Wiley & Sons, 2010.

[19] C. Howson, *Successful Business Intelligence: Unlock the Value of BI & Big Data*. McGraw-Hill, 2013.

[20] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 6th ed. Pearson, 2017.

[21] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, 2018.

[22] A. Raza and K. Kilic, "Predictive analytics in agriculture: Forecasting prices of pulses," *International Journal of Forecasting*, vol. 35, no. 3, pp. 1155-1166, 2019.

[23] R. E. Clark, *Developing Technical Training: A Structured Approach for Developing Classroom and Computer-based Instructional Materials*. London: Addison-Wesley, 2010.

[24] A. Zhang, H. Sun, D. Woodruff, and Y. Wang, "Auto-ML: From data to deployment," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 3225-3228, 2020, DOI: 10.14778/3415478.3415481.