**JURIKOM (Jurnal Riset Komputer)**, Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9584
Hal 763-768
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

# Predict Goods Demand Using the XGBoost Method Based on Sales Historical Data

**Badriah Nursakinah*, Nurhalimah, Yuda Samudra**

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1,*]dosen02779@unpam.ac.id, [2]dosen02956@unpam.ac.id, [3]dosen02623@unpam.ac.id
Email Penulis Korespondensi: dosen02779@unpam.ac.id
Submitted **09-02-2026;** Accepted **30-04-2026;** Published **30-04-2026**

**Abstract**

Predicting the demand for goods is an important aspect of inventory management and operational planning because inaccurate predictions can lead to overstock or shortages of goods. This study aims to predict the demand for goods using the Extreme Gradient Boosting (XGBoost) algorithm based on historical sales data. The dataset used contains information on the transaction date, number of sales, stock, price, and time index, which is then processed through the preprocessing and feature engineering stages, including the formation of temporal features and sales lag features. Data sharing is carried out using a time series split approach to maintain the chronological order of the data. The XGBoost model is optimized using GridSearchCV with the TimeSeriesSplit validation scheme. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Symmetric Mean Absolute Percentage Error (SMAPE). The results showed that the model produced an MAE score of 54.13 and an RMSE of 77.60, while a SMAPE score of 43.13% showed an acceptable relative error rate in highly fluctuating sales data. Feature importance analysis shows that previous period (lag_1) sales and weekly patterns are the most dominant factors in demand predictions. These results prove that XGBoost is effectively used for historical data-driven demand prediction of goods and has the potential to support inventory management decision-making.

**Keywords**: Sales Historical Data; Machine Learning; Demand Prediction; XGBoost; Time Series

## 1. INTRODUCTION

The increasingly competitive development of the business and trade world requires companies to have an accurate inventory planning and control system. One of the important aspects of inventory management is the ability to accurately predict the demand for goods. Inaccuracies in demand forecasting can lead to serious problems, such as overstocking that increases storage costs or stockouts that result in lost sales opportunities and decreased customer satisfaction [1]. Therefore, the prediction of demand for goods is a crucial issue that needs to be handled systematically and based on data.

Demand for goods is generally volatile and influenced by various factors, such as historical sales patterns, price changes, stock availability, seasonal factors, and consumer behavior. The complexity of these factors causes conventional methods, such as simple statistical approaches, to often be less able to capture the nonlinear patterns and long-term dynamics of sales data. As a result, the results of the predictions produced are not optimal and less adaptive to rapid changes in demand patterns [2]. Along with the development of technology, machine learning-based approaches are starting to be widely used to solve the problem of demand prediction. This method has the advantage of modeling complex and nonlinear relationships in time series data [3]. One of the machine learning algorithms that is widely used for prediction is Extreme Gradient Boosting (XGBoost), which is a gradient boosting-based ensemble learning method with high performance and good computing efficiency. XGBoost is able to combine a number of decision trees to iteratively minimize prediction errors [4][5].

Several previous studies have shown that machine learning-based methods provide better results than traditional statistical methods in the context of demand prediction. This study reports that the ensemble model is able to improve the accuracy of retail demand prediction compared to conventional approaches [6]. Furthermore, Kumar and Patel proved that XGBoost has superior performance over Random Forest and Support Vector Regression in sales predictions [7]. These results show the great potential of XGBoost in addressing demand forecasting problems.

Other studies have also integrated time series features with gradient boosting algorithms to predict seasonal demand and obtain fairly accurate results, but the evaluation is still limited to basic error metrics without model stability analysis [8]. Subsequent research also applied machine learning algorithms to predict product demand, but did not examine the influence of historical features in depth through feature importance analysis[9]. While the research conducted by Sami emphasizes the importance of model interpretability through feature importance analysis in XGBoost, it has not yet been combined with error distribution analysis [10].

Based on the study of related studies in the last five years, it can be identified that there is a research gap. Most studies still focus on improving prediction accuracy alone, without integrating comprehensive performance evaluation and model stability analysis. In addition, there is still limited research that integrates hyperparameter optimization, the use of various evaluation metrics (MAE, RMSE, MAPE, and SMAPE), and the analysis of error distribution and feature importance in a single commodity demand prediction framework.

Therefore, research is needed that not only focuses on prediction accuracy but is also able to provide a deeper understanding of model behavior and the factors that influence prediction outcomes. This approach is expected to produce a prediction model that is not only accurate but also stable and well-interpretable, making it more relevant to apply in the context of inventory management decision-making. Based on this background, this study aims to build a model of product

**JURIKOM (Jurnal Riset Komputer)**, Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9584
Hal 763-768
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

demand prediction using the XGBoost algorithm based on historical sales data with a time series approach. This research is expected to be able to produce a prediction model with good performance, evaluated using comprehensive error metrics, and provide an in-depth analysis through error distribution and feature importance. Thus, the results of this study are expected to make a practical and academic contribution to the development of a more accurate and reliable demand prediction system for goods.

# 2. RESEARCH METHODOLOGY

## 2.1 Research Stages

This research is carried out through several stages that are systematically arranged to ensure that the research process runs in a structured manner and the results obtained are in accordance with the research objectives. The research stages start from data collection to evaluation and analysis of the predicted results of demand for goods. The stages of the research are as follows:

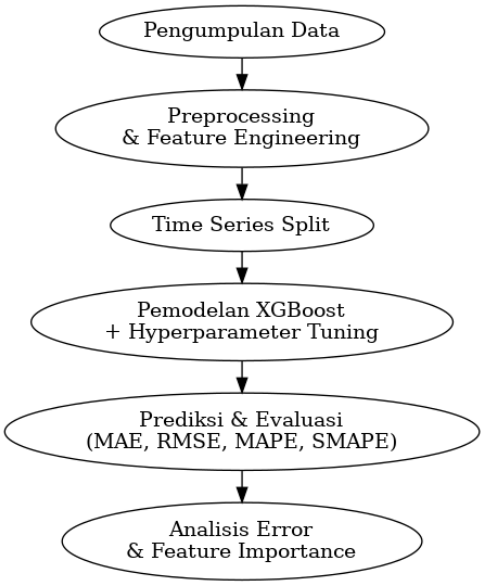


**Figure 1.** Research Stages

The explanation of the research stages is in Figure 1. It is as follows:

a. Data Collection

The initial stage of the research begins with the collection of historical sales datasets that are used as a basis for predicting the demand for goods. The dataset used was taken from the Kaggle of retail sales forecasting data. The dataset used contains several main attributes, namely the date of the transaction (data), the number of sales (venda) as the target variable, the stock of goods (estoque), the price of the product (preco), and the time index. This data represents the pattern of demand for goods over a given period of time and reflects sales fluctuations affected by temporal, price, and stock availability factors.

b. Preprocessing and Feature Engineering

In the preprocessing stage, the date attribute is converted to the appropriate time format, and the data is sorted in chronological order to maintain time series consistency. Furthermore, feature engineering is carried out to improve the model's ability to capture demand patterns. Temporal features such as year, month, day, and day of the week are extracted from the date attribute. In addition, a sales lag feature (lag-1, lag-7, and lag-14) was formed, which represents the dependence of current demand on sales in the previous period. This process aims to enrich historical information relevant to the prediction model [11][12].

c. Time Series Split

After the data goes through the preprocessing and feature engineering stages, the dataset is divided into training data and test data using the split time series approach [13]. This division is carried out in chronological order, where the training data comes from the initial period and the test data comes from the next time period. This approach is used to simulate real prediction conditions, i.e., predict demand in future periods based on previous historical data, as well as avoid information leakage [14].

d. XGBoost Modeling and Hyperparameter Tuning

**JURIKOM (Jurnal Riset Komputer)**, Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9584
Hal 763-768
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

The modeling stage was carried out by applying the Extreme Gradient Boosting (XGBoost) algorithm as a regression model to predict the amount of demand for goods. The model is trained using training data by utilizing all the features that have been formed [15]. To improve model performance, hyperparameter optimization is performed using GridSearchCV with the TimeSeriesSplit validation scheme. The optimized parameters include the number of trees, tree depth, learning rate, and subsampling ratio. This stage aims to obtain the model configuration that produces the lowest prediction error [16][17].

e. Prediction and Evaluation

The best model of the optimization results is then used to make demand predictions on the test data. The performance of the model was evaluated using several metrics, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Symmetric Mean Absolute Percentage Error (SMAPE). The use of several evaluation metrics aims to provide a comprehensive picture of the level of prediction errors, both absolute and relative, especially in sales data that has high fluctuations [18][19].

f. Error Analysis dan feature Importamce

The final stage of the research is error analysis and feature importance. Error analysis was carried out through visualization of residual distribution, the relationship of error to predictive value, and changes in absolute error over time to assess the stability and potential bias of the model. Furthermore, feature importance analysis is used to identify the most influential variables in the demand prediction process [20]. The results of this analysis show the contribution of each feature, in particular historical features of sales and temporal features, so that the built model is not only accurate but also can be interpreted well [21].

# 3. RESULTS AND DISCUSSION

## 3.1 Demand Prediction Modeling Results

In the modeling stage, the Extreme Gradient Boosting (XGBoost) algorithm was applied to predict the demand for goods based on historical sales data that had gone through the preprocessing and feature engineering process. The model was trained using training data that was arranged chronologically and tested on test data to maintain time series consistency. The prediction results show that the XGBoost model is able to follow the general patterns and fluctuating trends of the demand for goods. In low- to medium-demand periods, the predicted value is relatively close to the actual value, which indicates the model's ability to capture historical patterns of sales. However, in some periods with very high spikes in demand, models still tend to produce predictions that are lower than actual values, which shows the limitations of the model in predicting extreme values.

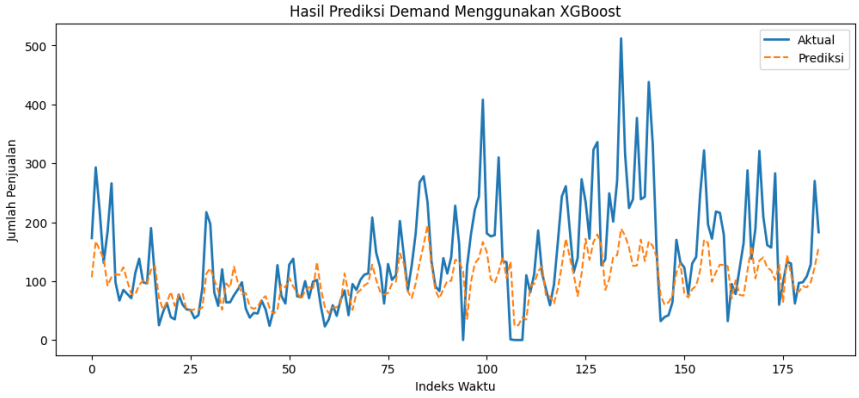


**Figure 2**. Prediction Results Using XGBoost

## 3.2 Model Performance Evaluation

The evaluation of model performance was carried out using several metrics, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Symmetric Mean Absolute Percentage Error (SMAPE). The test results showed that the model produced an MAE value of 54.13 and an RMSE of 77.60, which indicates that, on average, the difference between the predicted value and the actual value is still at an acceptable level. The SMAPE value of 43.13% indicates that the model's relative error rate is in the medium category, which is still relevant for high-fluctuating sales data. Meanwhile, the very large value of MAPE is influenced by the actual sales value that is very small or close to zero, distorting the error percentage calculation. Therefore, SMAPE is considered more representative than MAPE in the context of this study.

**JURIKOM (Jurnal Riset Komputer)**, Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9584
Hal 763-768
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

```
MAE   : 54.128055572509766
RMSE  : 77.59796920164051
MAPE  : 1105103518618.9363
SMAPE : 43.129099338667004
```

**Figure 3.** Model Performance Evaluation Results

### 3.3 Distribution Analysis and Error Patterns

The Error distribution shows the spread of the difference between the actual value and the predicted value of the demand for goods generated by the XGBoost model. Based on the histogram, the majority of error values are concentrated around zeros, which indicates that the model does not have a significant systematic bias in making predictions. This shows that, in general, the results of the predictions produced are quite close to the actual value. However, there are still some error values that spread to the positive and negative sides with a fairly wide range, which is indicated by the distribution tails on both sides of the histogram. The existence of these large errors indicates that there are certain periods where the model has difficulty predicting demand accurately, especially under extreme or unstable demand conditions. Overall, the relatively symmetrical and centralized error distribution pattern around zero suggests that the XGBoost model has a fairly stable performance, although it still has limitations in handling significant demand spikes.

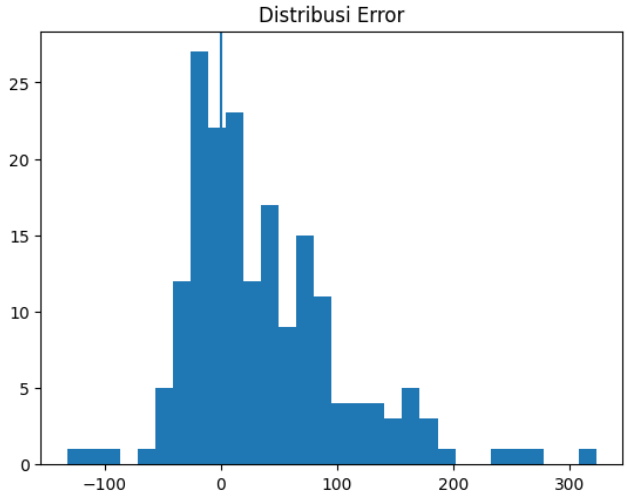


**Figure 4.** Error Distribution

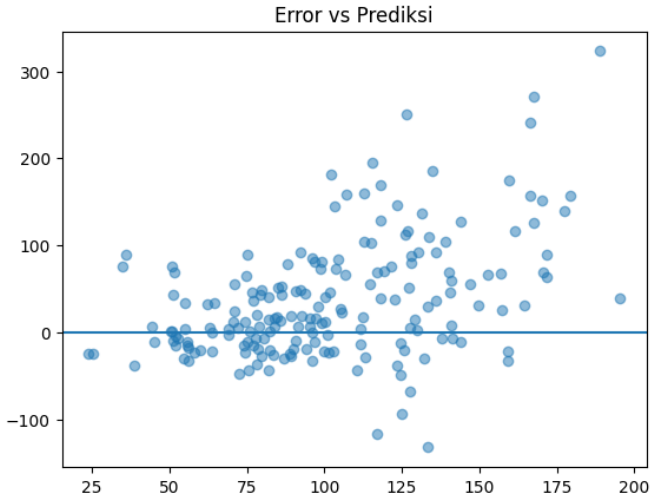


**Figure 5.** Error vs Prediction Results

Figure 5. shows the relationship between the predicted value of the demand for goods generated by the XGBoost model and the error value (the difference between the actual and predicted values). The dots on the graph are scattered around the zero line, which suggests that, in general, the model does not produce consistent prediction bias, either in the form of overestimation *or* underestimation tendencies. The relatively random distribution of points around the zero line indicates that the model is able to provide fairly stable predictions at various levels of demand. However, it can be seen that error variance tends to increase at higher prediction values. In the low to medium prediction range, the error is

**JURIKOM (Jurnal Riset Komputer)**, Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9584
Hal 763-768
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

relatively small and more concentrated, while at the high prediction value, the error spread becomes wider and several extreme values appear. This pattern indicates the presence of a symptom of heteroscedasticity, where predictive uncertainty increases as the demand for goods increases. These conditions indicate that the XGBoost model is more accurate in predicting normal demand, but still has limitations in handling large demand spikes. Overall, this graph shows that although the model has stable performance, improved accuracy at high demand is still a further development opportunity.

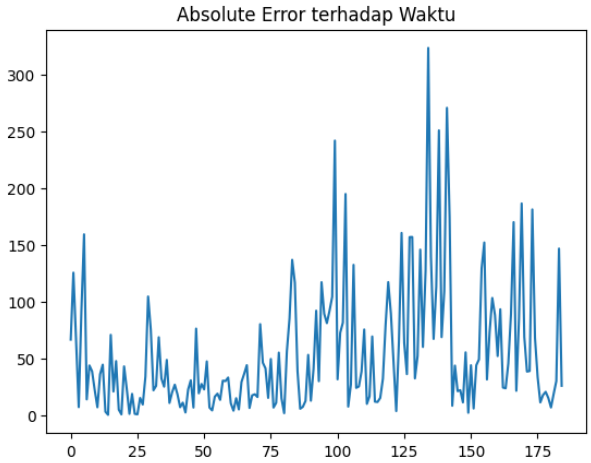


**Figure 6.** Absolute Error Against Time

Figure 6 shows the large change in the predictive errors of the goods demand generated by the XGBoost model over each time period. From the graph, it can be seen that in most of the early to medium period, the absolute error value is relatively low and stable, which indicates that the model can predict demand fairly accurately under normal conditions. However, in certain periods, there is a significant spike in absolute errors, especially in the middle to the end of the time series. These spikes indicate the presence of anomalous conditions or changes in demand patterns that the model cannot fully capture, such as a sudden increase in demand or extreme sales fluctuations. Overall, this graph indicates that although the XGBoost model has stable performance most of the time, the accuracy of the predictions tends to decline in periods of high demand dynamics, so further development is needed to improve the model's resilience to extreme changes in demand patterns.

### 3.4 Feature Importance Analysis

Feature importance analysis was carried out to evaluate the contribution level of each input variable to the prediction results generated by the XGBoost model. Where the results of feature importance can be seen as shown in Figure 7.

| | Feature | Importance |
|---|---|---|
| 6 | lag_1 | 0.382487 |
| 5 | dayofweek | 0.149786 |
| 0 | estoque | 0.086746 |
| 7 | lag_7 | 0.080344 |
| 3 | month | 0.062309 |
| 2 | year | 0.062163 |
| 8 | lag_14 | 0.060849 |
| 4 | day | 0.059434 |
| 1 | preco | 0.055881 |

**Figure 7.** Feature Importance

Figure 7 is The results of the feature importance analysis show that historical features of sales are the most dominant factor in predicting demand for goods. The lag_1 feature has the biggest contribution, which shows that sales in the previous period are the main indicator in predicting current demand. In addition, the lag_7 and dayofweek features also made significant contributions, indicating the presence of a weekly seasonal pattern in sales data. The stock (estoque) and price (preco) variables have less influence than historical features of sales, but they still play a role in helping models adjust predictions to certain conditions. These findings confirm that reliance on historical sales patterns is a key component in XGBoost-based demand prediction models.

**JURIKOM (Jurnal Riset Komputer)**, Vol. 13 No. 2, April 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i2.9584
Hal 763-768
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

# 4. CONCLUSION

This research has successfully implemented the Extreme Gradient Boosting (XGBoost) algorithm to predict the demand for goods based on historical sales data. The results of the study show that the machine learning-based approach can overcome the problem of inaccuracy of demand prediction that often occurs in conventional methods, especially in sales data that is volatile. Through preprocessing, time series-based feature engineering, and hyperparameter optimization using GridSearchCV, the XGBoost model is able to capture historical and seasonal patterns in sales data quite well. Evaluation of the model's performance using MAE, RMSE, MAPE, and SMAPE showed that the model had an adequate level of accuracy, with a MAE value of 54.13 and an RMSE of 77.60, while a SMAPE value of 43.13% indicated that the model's relative error was still within acceptable limits. Analysis of the error distribution showed that the model lacked significant systematic bias, although limitations were still found in predicting extreme demand. In addition, the results of the feature importance analysis confirm that sales in the previous period were the most dominant factor in determining current demand. Overall, this study proves that XGBoost is effectively used as a historical data-driven product demand prediction solution and has the potential to be a tool to support more accurate decision-making in inventory management.

# REFERENCES

[1] D. N. Gono, H. Napitupulu, and Firdaniza, "Silver Price Forecasting Using Extreme Gradient Boosting (XGBoost) Method," *Mathematics*, vol. 11, no. 18, 2023, doi: 10.3390/math11183813.
[2] S. Malik, M. Khan, M. K. Abid, and N. Aslam, "Sales Forecasting Using Machine Learning Algorithm in the Retail Sector," vol. 06, no. 02, 2024.
[3] J. Garcke and R. Roscher, "Explainable Machine Learning," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 1, pp. 169–170, 2023, doi: 10.3390/make5010010.
[4] A. K. Konyal\io\uglu, T. B. Apayd\in, \.Ilhan Turhan, A. Soydal, and T. Özcan, "An Extreme Gradient Boosting Model Optimized with Genetic Algorithm for Sales Forecasting of Retail Stores," in *Industrial Engineering in the Industry 4.0 Era*, N. M. Durakbasa and M. G. Gençy\ilmaz, Eds., Cham: Springer Nature Switzerland, 2024, pp. 59–67.
[5] H. Xie, S. Chen, C. Lai, G. Ma, and W. Huang, "Forecasting the clearing price in the day-ahead spot market using eXtreme Gradient Boosting," *Electr. Eng.*, vol. 104, no. 3, pp. 1607–1621, 2022, doi: 10.1007/s00202-021-01410-6.
[6] A. Panarese, G. Settanni, V. Vitti, and A. Galiano, "Developing and Preliminary Testing of a Machine Learning-Based Platform for Sales Forecasting Using a Gradient Boosting Approach," *Appl. Sci.*, vol. 12, no. 21, 2022, doi: 10.3390/app122111054.
[7] A. K. Sharma, L.-H. Li, and R. Ahmad, "Default Risk Prediction Using Random Forest and XGBoosting Classifier BT - 2021 International Conference on Security and Information Technologies with AI, Internet Computing and Big-data Applications," G. A. Tsihrintzis, S.-J. Wang, and I.-C. Lin, Eds., Cham: Springer International Publishing, 2023, pp. 91–101.
[8] P. H. Vuong, T. T. Dat, T. K. Mai, P. H. Uyen, and P. T. Bao, "Stock-Price Forecasting Based on XGBoost and LSTM," 2022, doi: 10.32604/csse.2022.017685.
[9] M. Dehvari, S. Farzaneh, and E. Forootan, "Forecasting rainfall events based on zenith wet delay time series utilizing eXtreme gradient boosting (XGBoost)," *Adv. Sp. Res.*, vol. 75, no. 3, pp. 2584–2598, 2025, doi: https://doi.org/10.1016/j.asr.2024.11.013.
[10] S. Ben Jabeur, S. Mefteh-Wali, and J. L. Viviani, "Forecasting gold price with the XGBoost algorithm and SHAP interaction values," *Ann. Oper. Res.*, vol. 334, no. 1–3, 2024, doi: 10.1007/s10479-021-04187-w.
[11] W. A. Nugroho, F. D. Rachman, B. K. Siahu, I. A. Iswanto, and S. Joddy, "Hybrid Ensemble Model Approaches for Stock Price Forecasting Using LSTM, Random Forest, ARIMA, and Linear Regression as Meta-Learner," *Procedia Comput. Sci.*, vol. 269, pp. 901–910, 2025, doi: https://doi.org/10.1016/j.procs.2025.09.033.
[12] N. Rofiq and S. L. M. Sitio, *Pengenalan Dasar Analisis Data dengan Python di Google Colab*. Eureka Media Aksara, 2024.
[13] R. Natras, B. Soja, and M. Schmidt, "Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting," *Remote Sens.*, vol. 14, no. 15, 2022, doi: 10.3390/rs14153547.
[14] G. Behera, A. Bhoi, and A. K. Bhoi, "A Comparative Analysis of Weekly Sales Forecasting Using Regression Techniques," in *Intelligent Systems*, S. K. Udgata, S. Sethi, and X.-Z. Gao, Eds., Singapore: Springer Nature Singapore, 2022, pp. 31–43.
[15] Jan Melvin Ayu Soraya Dachi and Pardomuan Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *J. Ris. Rumpun Mat. Dan Ilmu Pengetah. Alam*, vol. 2, no. 2, pp. 87–103, 2023, doi: 10.55606/jurrimipa.v2i2.1470.
[16] S. L. M. Sitio *et al.*, "Comparison of the Ensemble Xgboost and Transformer Models With Machine Learning for Classification of Indonesian Music Moods of the 70'S and 80'S Era," *J. Theor. Appl. Inf. Technol.*, vol. 102, no. 24, pp. 9157–9165, 2024.
[17] M. Mohamed, F. E. Mahmood, M. A. Abd, M. Rezkallah, A. Hamadi, and A. Chandra, "Load Demand Forecasting Using eXtreme Gradient Boosting (XGboost)," in *2023 IEEE Industry Applications Society Annual Meeting (IAS)*, 2023, pp. 1–7. doi: 10.1109/IAS54024.2023.10406613.
[18] K. K. Yun, S. W. Yoon, and D. Won, "Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process," *Expert Syst. Appl.*, vol. 186, p. 115716, 2021, doi: https://doi.org/10.1016/j.eswa.2021.115716.
[19] M. Hisyam, Z. Fitri, H. Al, and K. Aidilof, "Implementasi Algoritma XGBoost dengan Walk Forward Validation untuk Prediksi Harga Emas Antam," vol. 12, no. 4, pp. 403–413, 2025, doi: 10.30865/jurikom.v12i4.8693.
[20] A. Mitra, A. Jain, A. Kishore, and P. Kumar, "A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach," *Oper. Res. Forum*, vol. 3, no. 4, p. 58, 2022, doi: 10.1007/s43069-022-00166-4.
[21] S. Sharma and P. Chaudhary, "Machine learning and deep learning," *Quantum Comput. Artif. Intell. Train. Mach. Deep Learn. Algorithms Quantum Comput.*, pp. 71–84, 2023, doi: 10.1515/9783110791402-004.