Scientific Journal of Computer Science

futuristech
PT. Teknologi Futuristik Indonesia

Date of publication February 19, 2026, date of current version February 19, 2026.
Digital Object Identifier 10.64539/sjcs.v2i1.2026.407

Article

Performance Evaluation of Tree-Based Machine Learning Algorithms for Medical Relief Supply Demand Forecasting

Roman Bariring Villones[1,*], Janette Templonuevo Vargas[1]

[1] Graduate School Department, La Consolacion University Philippines, Malolos, 3000 Bulacan, Philippines; roman.villones@email.lcup.edu.ph; janette.vargas@email.lcup.edu.ph
* Correspondence

The authors declare that no funding was received for this study.

Abstract: Accurate demand forecasting is critical in medical relief supply chains where prediction errors can lead to stockouts, delayed response, or inefficient allocation of limited resources. While machine learning (ML) approaches have demonstrated superior predictive capabilities compared to traditional statistical methods and existing research often treats ML as a homogeneous category and rarely conducts systematic benchmarking within specific algorithm families. Furthermore, many studies rely on default model configurations that limiting the reproducibility and failing to fully assess its robustness under volatile demand conditions common in humanitarian logistics. This study addresses these gaps by systematically evaluating multiple tree-based machine learning algorithms for medical relief supply demand forecasting under a structured framework. The research integrates GridSearchCV as hyperparameter optimization, repeated K-fold cross-validation, and statistical significance testing to ensure fair comparison and robustness assessment. The findings indicate that advanced gradient boosting models outperform single-tree and simpler ensemble approaches in terms of predictive accuracy and stability. CatBoost consistently achieved the lowest prediction errors, the narrowest residual dispersion, and the most stable cross-validation performance. Although statistically comparable to other advanced boosting frameworks, CatBoost demonstrated superior robustness during volatile demand conditions and demand surges. These results provide both methodological and practical contributions by establishing a benchmarking framework and identifying a stable forecasting model that suitable for operational deployment in AI-driven medical relief inventory system.

Keywords: Demand forecasting; Medical relief supply chain; Tree-based machine learning; Hyperparameter optimization; CRISP-DM.

1. Introduction

Demand forecasting plays a fundamental role in supply chain performance as prediction accuracy that directly affects the inventory control, logistics coordination, and service level attainment across industries [1]–[3]. In healthcare and humanitarian logistics, forecasting accuracy becomes even more critical because errors may result in stockouts and delayed responses. Traditional statistical forecasting approaches often struggle to capture nonlinear demand behavior and complex feature interactions present in real-world supply chain data [1], [2]. Consequently, machine learning (ML) techniques have gained increasing attention due to their ability to model nonlinearities and high-dimensional relationships more effectively [1], [2], [4], [5]. Among these approaches, a tree-based ML algorithm have frequently demonstrated superior predictive performance compared to classical statistical models, particularly when handling diverse and complex datasets [1], [2], [6]. However, their effectiveness remains highly dependent on data quality, feature engineering, and hyperparameter configuration, especially under volatile demand conditions [1], [7].

Despite the growing body of literature demonstrating the potential of ML-based forecasting, several important gaps remain unresolved. First, many studies treat machine learning as a broad category rather than system-

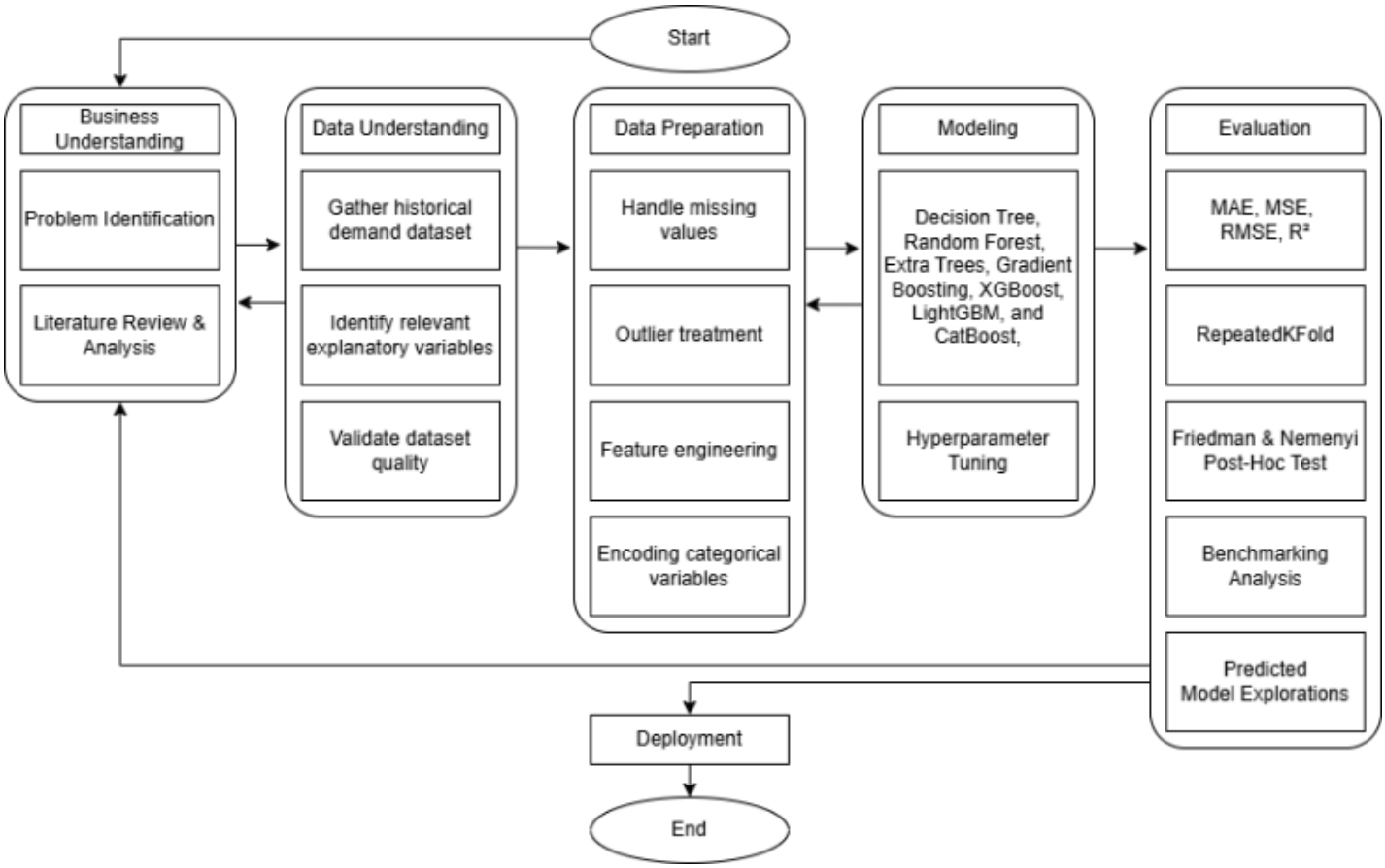

Figure 1. Cross-Industry Standard Process for Data Mining (CRISP-DM) framework.

atically examining the performance differences among specific algorithm families, particularly tree-based models [8], [9]. As a result, there is limited empirical evidence that rigorously isolates and evaluates the comparative effectiveness of tree-based algorithms within medical relief supply demand forecasting contexts. Second, although prior studies acknowledge that model performance is highly sensitive to hyperparameter selection, many rely on default configurations or limited tuning procedures [6], [10]. While optimization techniques such as GridSearch are frequently referenced, a few studies rigorously integrate the structured hyperparameter optimization into a comparative benchmarking framework particularly in high-risk and irregular demand environments [4], [11]. This limits the reproducibility, robustness, and practical applicability. Third, existing forecasting studies often emphasize aggregate accuracy metrics without adequately evaluating robustness, generalization stability, and statistical significance across validation folds. In volatile operational environments, these additional dimensions of evaluation are essential for reliable deployment.

The persistence of this problem can be attributed to both methodological and contextual limitations in prior research. Comparative forecasting studies frequently focus on broad comparisons between traditional statistical models and ML approaches, rather than conducting controlled benchmarking within specific ML algorithm families [12], [13]. Additionally, although optimization is widely recognized as crucial for performance enhancement, it is often under-exploited in empirical evaluations [13], [14]. Moreover, the medical relief and humanitarian demand data are inherently irregular and high-risk in making a robust validation that more complex. The lack of structured and reproducible evaluation frameworks tailored to such volatile environments has resulted in limited systematic analysis of algorithm stability and operational reliability [15], [16], [17]. Consequently, there remains a clear need for benchmarking of tree-based ML models under optimized and statistically validated conditions.

In response to these identified gaps, this study systematically evaluates multiple tree-based machine learning algorithms for medical relief supply demand forecasting within a structured and reproducible framework. Building upon evidence that ML approaches can outperform traditional forecasting methods in capturing nonlinear and complex demand patterns [4], [5], this research advances the field by integrating algorithm-specific benchmarking with systematic optimization and robustness analysis.

The contributions of this study are fourfold:

1) It provides a rigorous comparative evaluation of selected tree-based ML algorithms and offering

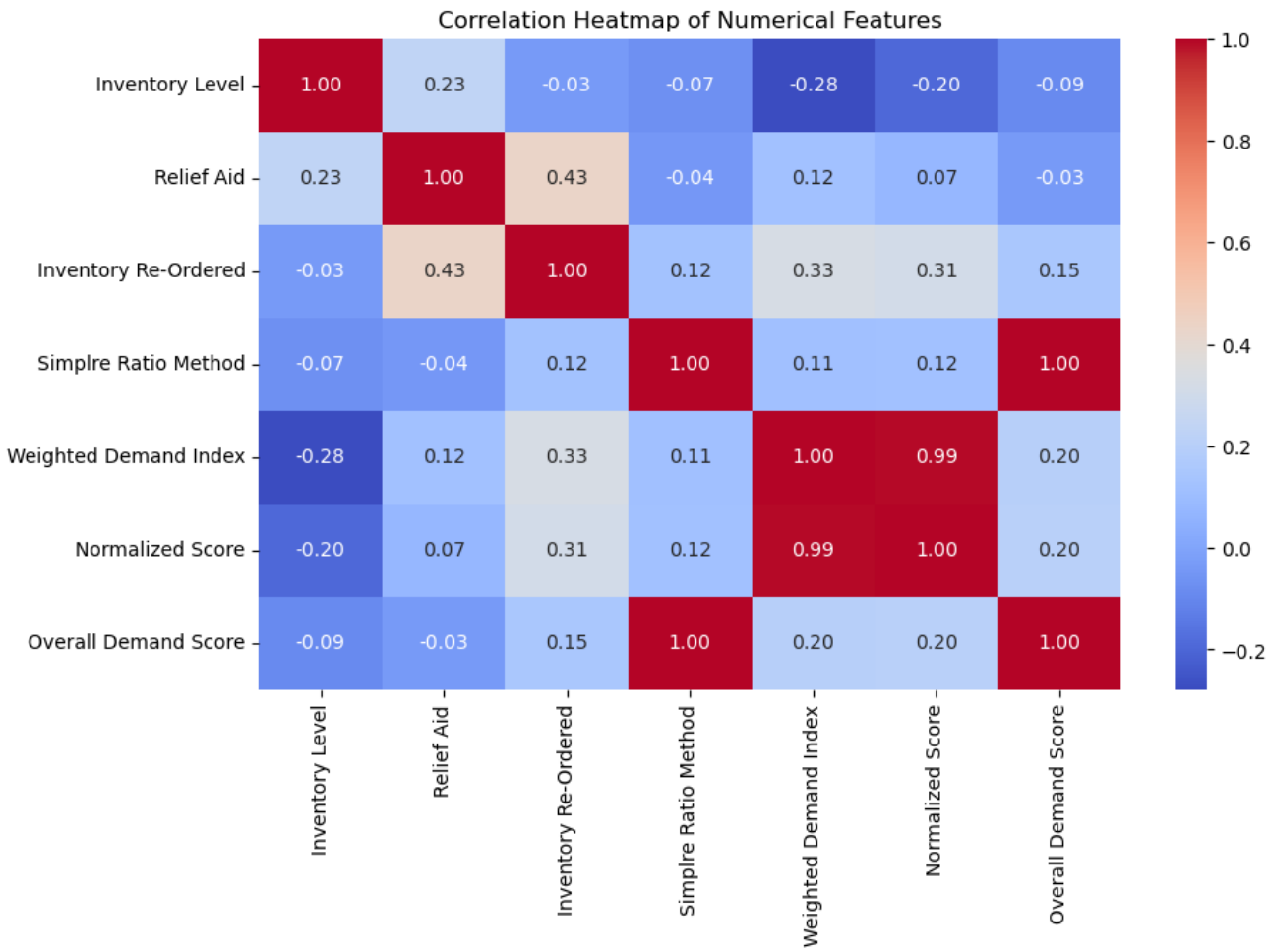

Figure 2. Correlation Heatmap of Numerical Features.

empirical evidence on their relative strengths and limitations in medical relief demand forecasting [12], [13].

2) The study integrates GridSearch-based tuning into the evaluation framework, demonstrating how systematic optimization enhances predictive accuracy, and generalization of performance in area that previously identified as under-utilized [13], [14].
3) Beyond conventional accuracy metrics, the research incorporates repeated cross-validation and statistical testing to assess model stability under volatile demand conditions and addressing concerns in high-risk forecasting environments [15], [16].
4) The findings generate actionable insights for healthcare and humanitarian supply chain managers by identifying models that balance predictive performance, robustness, and interpretability, thus supporting data-driven medical relief planning [15], [16], [17].

Through this structured optimization-driven and statistically validated approach, the study strengthens the methodology in demand forecasting research and enhances its practical applicability in critical medical relief supply chain operations.

2. Methodology

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to guide the systematic development, evaluation, and validation of machine learning models for medical relief supply demand forecasting. CRISP-DM is a well-established and widely adopted analytical framework that ensures methodological rigor, transparency, and reproducibility in data-driven research by structuring the entire modeling process into iterative and interrelated phases [15], [18]. The framework consists of six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment shows in the Figure 1.

2.1. Business Understanding

In this phase, the research objectives are defined based on the operational challenges associated with medical relief supply demand forecasting. The primary objective is to improve forecasting accuracy through the application and comparative evaluation of optimized tree-based machine learning algorithms. Performance indicators and evaluation metrics are identified to ensure alignment between predictive outcomes and practical decision-making requirements in healthcare and humanitarian supply chain contexts [19].

2.2. Data Understanding

The data understanding phase involves an initial exploration of historical medical relief supply demand records to assess their structure and completeness. The dataset contains more than 76,000 rows and variables that comprising both qualitative (categorical) and quantitative (numerical) types. Understanding the distinction between qualitative and quantitative variables ensures appropriate preprocessing and modeling strategies for accurate and robust predictions [20].

2.3. Data Preparation.

During data preparation, the dataset is cleaned and transformed to ensure suitability for machine learning modeling. The dataset contains no missing values. The target variable for forecasting is Relief Aid, while the categorical variables are converted into numerical representations using encoding techniques suitable for machine learning algorithms. Proper data preparation is critical, as prior studies emphasize that forecasting performance is highly dependent on the quality, relevance, and representation of input features [21].

Figure 2 represents the results of correlation heatmap and it indicate the limited multicollinearity among most variables. The strongest relationship is observed between the Weighted Demand Index and the Normalized Score (r = 0.99), indicating that these measures convey nearly identical information and should not be treated as independent predictors. A perfect positive correlation (r = 1.00) between the Simple Ratio Method and the Overall Demand Score further confirms a deterministic scoring relationship rather than an urgent pattern. Inventory Level shows weak to moderate negative correlations with demand indicators, particularly with the Weighted Demand Index (r = −0.28) and the Normalized Score (r = −0.20). The Relief Aid variable exhibits a moderate positive correlation with Inventory Re-Ordered (r = 0.43), indicating that relief distribution activities are linked to replenishment decisions, while Inventory Re-Ordered also demonstrates moderate positive correlations with the Weighted Demand Index (r = 0.33) and Normalized Score (r = 0.31), reflecting the influence of demand intensity on restocking behavior. Most remaining relationships are weak ($|r| < 0.20$), suggesting that the variables capture distinct operational dimensions and thereby supporting the analytical robustness of the dataset while underscoring the importance of excluding highly collinear derived measures in model development.

2.4. Modeling.

During the modeling phase focuses on the implementation and evaluation of tree-based machine learning algorithms, which are widely recognized for their predictive performance in demand forecasting tasks. The algorithms used in this study include Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM, and CatBoost, covering both single-tree and ensemble approaches [22]–[24]. The dataset is partitioned into training (80%) and testing (20%) subsets to ensure robust model evaluation and to prevent overfitting.

To enhance predictive performance and ensure comparability across models, GridSearchCV is employed for systematic hyperparameter optimization, while 5-fold cross-validation is applied during training to assess model stability and generalization across different data splits [25], [26]. Structured hyperparameter tuning has been shown to improve the accuracy, robustness, and interpretability of tree-based models, particularly in scenarios with complex and non-linear demand patterns [16], [21], [27]. This approach enables fair and reproducible evaluation of each algorithm and facilitating a comprehensive comparison of their suitability for forecasting medical relief supply demand.

2.5. Evaluation

Model performance is evaluated using standard regression metrics and commonly applied in demand forecasting. Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) are the metrics provided for complementary insights into a predictive accuracy, error magnitude, and its model goodness-of-fit by enabling an assessment of each algorithm [13], [17], [21].

To ensure robustness against random data-split variability, a 10-fold cross-validation procedure with multiple repetitions was implemented. Beyond reporting average error metrics, a statistical significance testing was conducted to verify whether observed performance differences among models were non-random. Specifically, the Friedman test, a non-parametric statistical test suitable for comparing multiple machine learning algorithms across repeated experiments, was applied to the cross-validation results. Upon detecting statistically significant differences ($\alpha = 0.05$), a Nemenyi post-hoc analysis was performed to identify pairwise model differences.

2.5.1. Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

Measures the average absolute difference between the actual and predicted values. It provides a straightforward interpretation of prediction error in the same unit as the target variable and is less sensitive to extreme outliers compared to squared-error metrics [28]–[30].

2.5.2. Mean Squared Error (MSE)

$$MSE = 1/n\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

Table 1. Evaluation Results of Tree-Based Algorithms.

ML Algorithm	MAE	MSE	RMSE	R^2	Rank
CatBoost	1.630896	7.408818	2.721914	0.996167	1
XGBoost	2.482483	17.733944	4.211169	0.990826	2
Extra Trees	1.576271	18.541259	4.305956	0.990408	3
LightGBM	2.749605	21.843258	4.673677	0.988700	4
Random Forest	1.813812	22.941765	4.789756	0.988132	5
Decision Tree	3.055362	60.489293	7.777486	0.968708	6
Gradient Boosting	6.753341	143.974124	11.998922	0.925519	7

Table 2. GridSearchCV Result using Negative RMSE scoring.

ML Algorithm	Best RMSE	Best Parameters	Rank
CatBoost	2.688430	{'depth': 6, 'iterations': 700, 'learning_rate': 0.2}	1
Gradient Boosting	3.041938	{'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 300}	2
LightGBM	3.053807	{'learning_rate': 0.1, 'n_estimators': 300, 'num_leaves': 63}	3
XGBoost	3.343500	{'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 300}	4
Extra Trees	4.231030	{'max_depth': 30, 'min_samples_split': 2, 'n_estimators': 100}	5
Random Forest	4.699428	{'max_depth': 30, 'min_samples_split': 2, 'n_estimators': 200}	6
Decision Tree	7.221813	{'max_depth': 30, 'min_samples_leaf': 2, 'min_samples_split': 2}	7

Computes the average of the squared differences between actual and predicted values. By squaring the errors, MSE penalizes larger deviations more heavily and making it particularly useful for identifying models with significant prediction errors [28], [30].

2.5.3. Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

Is the square root of MSE and represents the standard deviation of prediction errors. RMSE retains the same unit as the target variable while emphasizing large errors and making it suitable for evaluating forecasting performance in high-impact applications such as medical relief supply planning [28]–[30].

2.5.4. Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

It quantifies the proportion of variance in the dependent variable that is explained by the model. It indicates the goodness-of-fit of the model, where values closer to 1 suggest stronger explanatory power and better predictive performance [30].

2.5.5. Repeated KFold

It is an extension of standard k-fold cross-validation in which the data partitioning process is repeated multiple times with different random splits. By averaging results across multiple folds and repetitions, the method provides a more reliable estimate of a model's generalization capability and mitigates random data-split bias [31].

2.5.6. Friedman Test

The Friedman test is a non-parametric statistical test used to detect significant differences among multiple related models across repeated experimental runs. Unlike ANOVA, it does not assume normality of residuals, making it particularly suitable for comparing machine learning algorithms evaluated across cross-validation folds [32].

2.5.7. Nemenyi Post-Hoc Test

The Nemenyi post-hoc test is conducted following a significant Friedman test to determine which specific model pairs differ statistically. It compares average model ranks and identifies significant differences using a critical difference (CD) threshold [33].

3. Results

Table 1 presents the comparative performance of the evaluated machine learning algorithms. CatBoost achieved the best overall performance and ranked first with consistently low MAE (1.63), MSE (7.41), and RMSE (2.72), indicating a high prediction accuracy and minimal error variance. Its high R^2 value (0.9962) further confirms its strong explanatory power and ability to capture underlying demand patterns. XGBoost ranked second, demonstrating balanced performance across all metrics with moderate MAE (2.48), MSE (17.73), RMSE (4.21) and a high R^2 (0.9908). The Extra Trees model ranked third, recording the lowest MAE (1.58) among all models but slightly higher MSE (18.54) and RMSE (4.31) values, indi-

Table 3. 10-fold Repeated K-Fold Cross Validation Result.

ML Algorithm	RMSE Mean	RMSE Standard Deviation
CatBoost	2.5308	0.1092
Gradient Boosting	2.7693	0.2025
LightGBM	2.8373	0.1872
XGBoost	3.1716	0.1957
Extra Trees	3.8195	0.2915
Random Forest	4.1986	0.2657
Decision Tree	6.3529	0.3190

Table 4. Friedman's Test Result.

Friedman Test	Result
Statistic	293.39
p-value	< 0.001

Table 5. Nemenyi post-hoc Test Result.

Comparing 'Catboost' with	p-value	Significant (α = 0.05)?
Decision Tree	0.003	YES
Random Forest	< 0.001	YES
Extra Trees	< 0.001	YES
Gradient Boosting	< 0.001	YES
XGBoost	0.832	NO
LightGBM	0.055	NO

cating accurate point predictions with increased sensitivity to larger errors.

LightGBM and Random Forest, ranked fourth and fifth respectively, and it showed a comparable performance with MAE values below (3.0) and RMSE values under (5.0). Although their R^2 values remained high (above 0.98), the increased MSE and RMSE indicate wider error dispersion compared to the top-ranked models. In contrast, Decision Tree and Gradient Boosting models exhibited substantially higher error metrics. The Decision Tree model recorded elevated MAE (3.06) and RMSE (7.78), reflecting reduced predictive stability. Meanwhile, Gradient Boosting performed the weakest among all evaluated models, with the highest MAE (6.75), MSE (143.97), and RMSE (12.00), coupled with the lowest R^2 (0.9255), indicating limited suitability for accurate medical relief supply demand forecasting.

Table 2 presents the results of hyperparameter optimization conducted using GridSearchCV with `neg_root_mean_squared_error` (neg-RMSE) as the scoring criterion. Among all evaluated algorithms, CatBoost achieved the lowest optimized RMSE (2.69), ranking first. The optimal configuration (depth = 6, iterations = 700, learning_rate = 0.2) indicates that moderate tree depth combined with a higher number of boosting iterations enhances predictive precision while maintaining stability. Gradient Boosting ranked second with an optimized RMSE (3.04) where increased tree depth and estimator count (max_depth = 7, n_estimators = 300) contributed combined with a higher number of boosting iterations enhances predictive precision while maintaining stability. LightGBM followed closely in third place, achieving an RMSE (3.05) with a configuration emphasizing a balanced number of leaves and estimators, then highlighting its efficiency in capturing nonlinear demand relationships.

XGBoost ranked fourth with an RMSE (3.34), demonstrating competitive performance under similar depth and learning rate settings. Extra Trees and Random Forest ranked fifth and sixth, respectively, with RMSE (< 4.20), indicating increased prediction variance despite deeper trees. The Decision Tree model performed the weakest, recording the highest RMSE (7.22) which reflects its limited generalization capability even after hyperparameter tuning.

Table 3 presents the results of the 10-fold Repeated K-Fold Cross-Validation indicate clear performance differences among the evaluated models. CatBoost achieved the lowest mean RMSE (2.5308 ± 0.1092), demonstrating both superior predictive accuracy and high stability across folds. Gradient Boosting (2.7693 ± 0.2025) and LightGBM (2.8373 ± 0.1872) also exhibited competitive performance but with higher variability. In contrast, traditional tree-based models such as Decision Tree (6.3529 ± 0.3190) and Random Forest (4.1986 ± 0.2657) showed substantially higher prediction errors.

Table 4 shows the results of Friedman test and it reveals a statistically significant difference among the seven evaluated models with statistics of 293.39 and p-value of 2.13×10^{-60}. The extremely small p-value indicates that the observed performance differences are unlikely to be due to random data split variability. Therefore, the null hypothesis that all models perform equivalently is rejected. To determine which specific models, differ significantly, a Nemenyi post-hoc analysis was subsequently conducted.

Table 5 shows the Nemenyi post-hoc analysis revealed that CatBoost significantly outperformed Decision Tree, Random Forest, Extra Trees, and Gradient Boosting. However, there is no statistically significant difference was observed between CatBoost and XGBoost (p = 0.832) or LightGBM (p = 0.055). These findings indicate that while CatBoost demonstrates superior performance over traditional tree-based ensemble methods, its performance is statistically comparable to other advanced gradient boosting frameworks.

Figure 3 shows the Residual analysis indicates that CatBoost exhibits the most concentrated error distribution centered around zero, with comparatively shorter tails than XGBoost and LightGBM. This suggests that a lower prediction variance and reduced extreme estima-

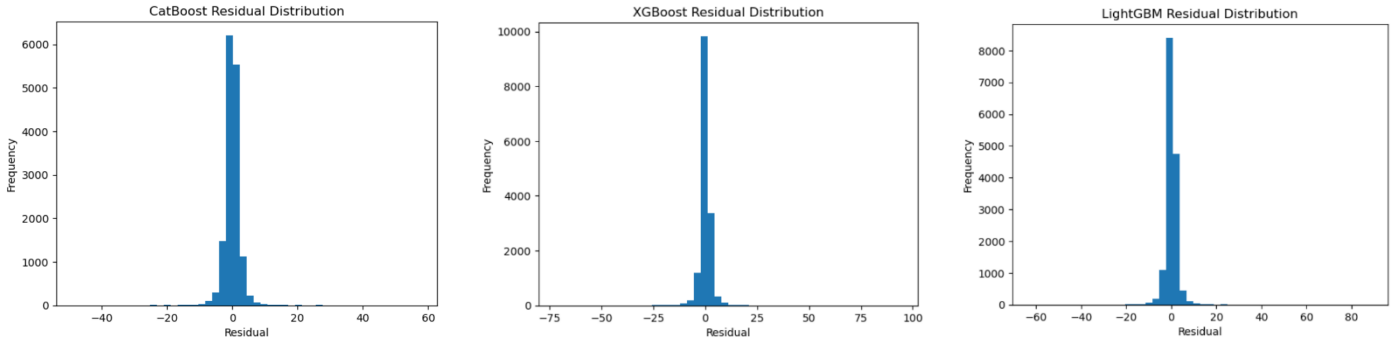

Figure 3. Residual Distribution Analysis.

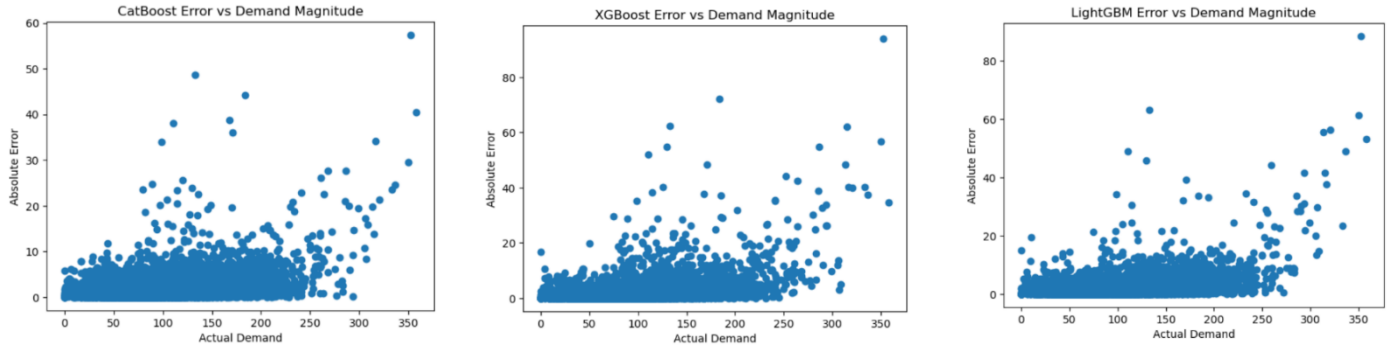

Figure 4. Demand Magnitude Analysis.

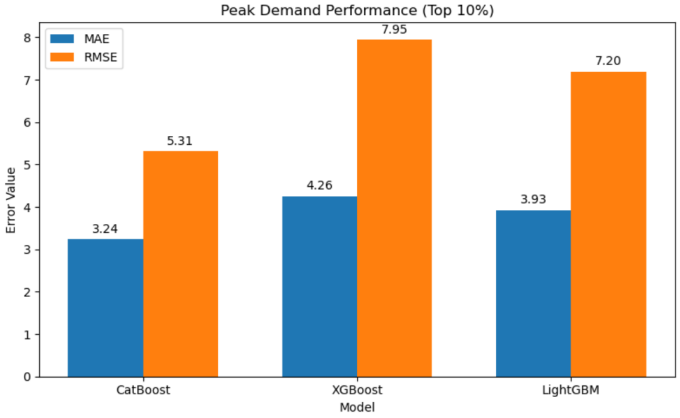

Figure 5. Peak demand performance analysis.

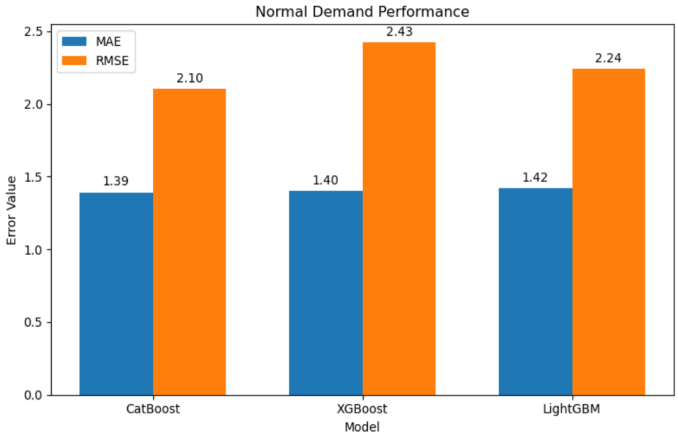

Figure 6. Normal demand performance analysis.

tion errors, particularly under fluctuating demand conditions, such stability is critical in military medical logistics where over- or under-estimation during emergency periods may lead to stockouts or resource wastage. Although all three models demonstrated low bias, CatBoost exhibited the narrowest residual dispersion and fewer extreme deviations and it indicates a superior robustness and stability during volatile demand conditions.

Figure 4 shows a demand magnitude analysis that indicates that the CatBoost maintains the most concentrated absolute error distribution across increasing demand levels with comparatively shorter extreme deviations than XGBoost and LightGBM. While all models exhibit increasing error variance at higher demand magnitudes, CatBoost shows a more controlled dispersion and fewer high-error outliers. In critical forecasting environments where demand surges are unpredictable and minimizing an extreme estimation error is essential to prevent stockouts or resource misallocation.

Figure 5 shows a Peak demand performance analysis (top 10% demand levels) indicates that CatBoost achieves the lowest MAE (3.24) and RMSE (5.31), demonstrating superior accuracy and reduced large-error deviations during high-demand periods. In comparison, XGBoost records the highest MAE (4.26) and RMSE (7.95), suggesting greater prediction variance and more pronounced extreme errors under peak conditions. LightGBM performs moderately (MAE: 3.93; RMSE: 7.20) but still exhibits higher dispersion than CatBoost.

Figure 6 shows a normal demand performance analysis that all three models achieve relatively similar accuracy under typical demand conditions. Cat-Boost records

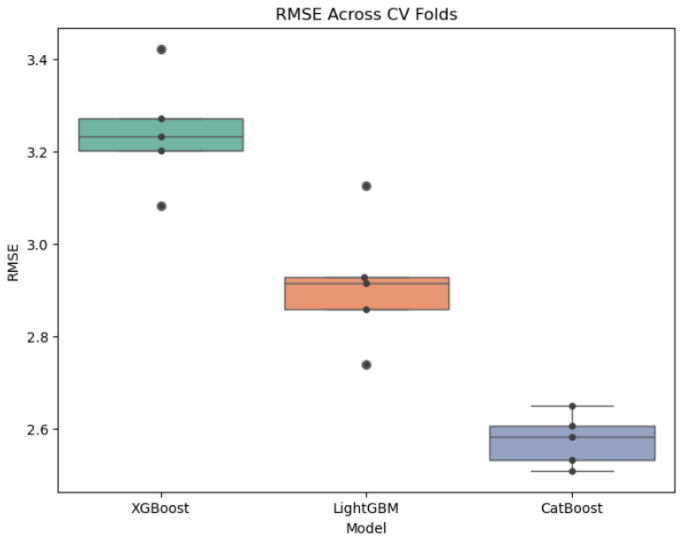

Figure 6. Cross-validation RMSE Analysis.

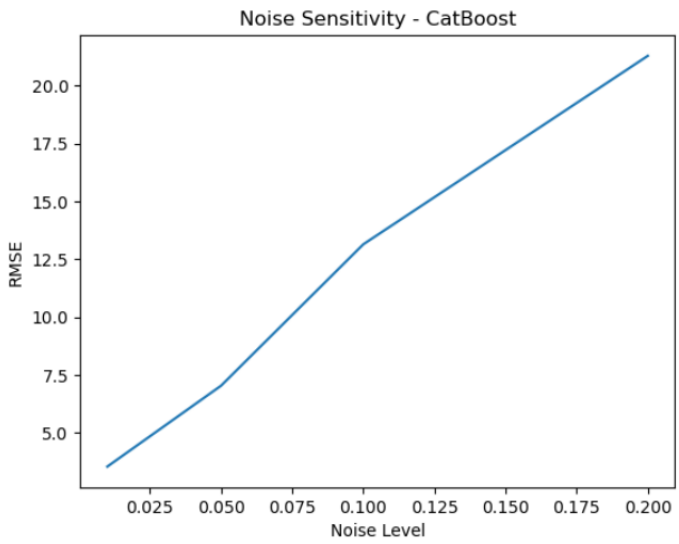

Figure 9. Noise Sensitivity Analysis.

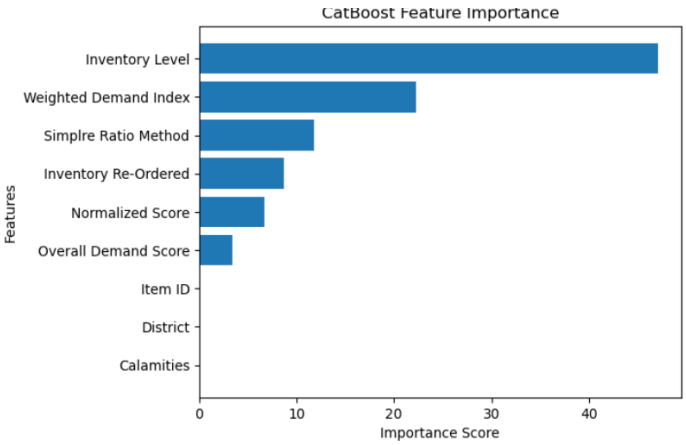

Figure 7. Feature Importance Analysis.

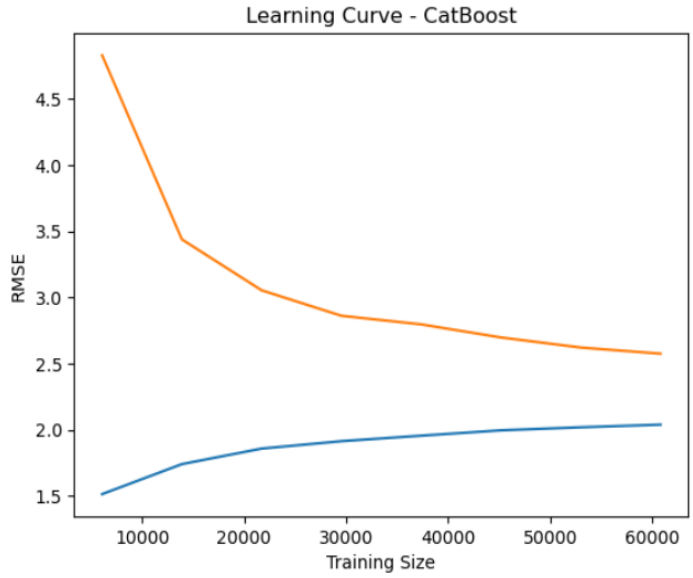

Figure 8. Learning Curve Analysis.

the lowest MAE (1.39) and RMSE (2.10), indicating slightly better precision and lower error variance. XGBoost demonstrates comparable MAE (1.40) but a higher RMSE (2.43), suggesting more occasional larger deviations. LightGBM performs marginally behind CatBoost (MAE: 1.42; RMSE: 2.24) with moderate dispersion.

Figure 6 shows a Cross-validation RMSE analysis that shows noticeable performance differences among the three models across folds. CatBoost achieves the lowest average RMSE (approximately 2.55–2.60), indicating better overall predictive accuracy and lower error magnitude. Its narrow interquartile range suggests strong consistency and minimal variability between validation folds. LightGBM records intermediate RMSE values (around 2.85–2.95), performing better than XGBoost but with slightly greater dispersion than CatBoost. While generally stable, the presence of a higher fold value indicates occasional increases in prediction error. XGBoost demonstrates the highest RMSE values (approximately 3.20–3.30), reflecting comparatively larger prediction errors. It also exhibits the widest spread and a noticeable higher outlier, suggesting greater variability and less stable performance across folds.

Figure 7 shows the feature importance analysis that indicates the Inventory Level (47.13%) is the most influential predictor, contributing nearly half of the model's decision-making weight. This is followed by the Weighted Demand Index (22.27%) and the Simple Ratio Method (11.79%), suggesting that demand-driven and stock-related indicators are the primary determinants of forecasting performance. Inventory Re-Ordered (8.68%), Normalized Score (6.73%), and Overall Demand Score (3.37%) provide moderate explanatory power. In contrast, Item ID, District, and Calamities exhibit negligible importance (each below 0.02%), indicating minimal direct contribution to prediction accuracy.

Figure 8 illustrate the learning curve analysis of CatBoost that demonstrates a stable generalization behavior as the training size increases. The validation RMSE decreases consistently with additional data and gradually stabilizes, indicating improved predictive capability and

convergence. Although a moderate gap between training and validation error is observed, the difference narrows as the dataset grows, suggesting controlled model complexity and absence of severe overfitting. These findings confirm that the model remains robust under varying data availability conditions, strengthening its suitability for real-world disaster-related demand forecasting.

Figure 9 shows the noise sensitivity analysis that indicates the CatBoost's RMSE increases progressively as noise levels rise. At minimal noise (0.01), the RMSE remains relatively low (3.55), demonstrating strong predictive stability. However, as noise increases to 0.05 and 0.1, RMSE grows substantially (7.04 and 13.15), reflecting reduced accuracy under data perturbation. At high noise levels (0.2), RMSE escalates significantly (21.29), indicating notable performance degradation. This pattern suggests that while CatBoost performs robustly under low-noise conditions, prediction accuracy becomes increasingly sensitive as data uncertainty intensifies.

4. Discussions

The findings of this study provide clear evidence that ensemble-based and optimized tree models substantially outperform single-tree and less robust ensemble approaches in forecasting medical relief supply demand. Machine learning ensemble methods such as CatBoost, XGBoost, and LightGBM consistently produced lower MAE and RMSE values, indicates more reliable and precise demand estimates. The comparison between baseline and tuned RMSE values further demonstrates that hyperparameter optimization yields the greatest benefits for boosting-based algorithms. Gradient Boosting, LightGBM, and CatBoost exhibited substantial reductions in RMSE following tuning, whereas simpler ensemble structures such as Random Forest and single Decision Tree models showed only marginal improvements.

After applying 10-fold Repeated K-Fold cross-validation, the results further reinforced model reliability. The low standard deviation observed for CatBoost indicates consistent predictive performance across different data partitions and demonstrating the robustness prior to statistical significance testing. To determine whether observed performance differences were statistically significant, Friedman's test and Nemenyi post-hoc analysis was subsequently conducted. The results indicate that while CatBoost demonstrates superior performance compared to traditional tree-based ensemble methods, its predictive performance is statistically comparable to other advanced gradient boosting frameworks such as XGBoost and LightGBM.

Residual analysis further supports this study, although the top three boosting models demonstrated similarly low bias, the CatBoost exhibited the narrowest residual dispersion and fewer extreme deviations. Under high-demand scenarios, CatBoost provides the most stable and reliable forecasts and minimizing a large positive or negative deviation that could result in stockouts and over-allocation of scarce medical resources. While differences among advanced boosting models are minimal during normal demand levels, CatBoost consistently maintains the most stable predictive performance.

Feature contribution analysis further reveals that the model relies predominantly on quantitative inventory and demand-related metrics rather than categorical or external contextual variables. This highlights the operational significance of stock levels, historical demand, and lead-time dynamics in shaping forecasting accuracy. Robustness testing under varying data availability conditions also confirms the model's stability. Learning curve evaluation demonstrates that CatBoost sustains strong generalization performance even when trained on reduced data subsets, strengthening its suitability for real-world disaster-related demand forecasting where historical data may be limited. However, noise sensitivity analysis suggests that while CatBoost performs robustly under low-noise conditions, prediction accuracy becomes increasingly sensitive as data uncertainty intensifies.

5. Conclusions and Recommendations

This study confirms that CatBoost demonstrates a superior stability, reliability, and robustness under volatile and uncertain demand environments. Although other advanced gradient boosting algorithms such as XGBoost and LightGBM achieved a statistically comparable levels of predictive accuracy. The CatBoost consistently exhibits a narrower error dispersion, it's stronger cross-validation stability, and more dependable performance during demand surges. These characteristics are particularly critical in medical relief supply forecasting, were minimizing large estimation errors directly contributes to preventing stockouts and maintaining operational readiness. The overall results indicate that CatBoost provides the most stable and operationally reliable forecasting performance for AI-driven medical relief inventory systems operating under dynamic and high-risk conditions.

Based on the findings of this study, the following recommendations are proposed:

1) CatBoost is recommended as the primary forecasting model for AI-driven medical relief inventory systems due to its demonstrated stability and robustness under volatile demand conditions.
2) Continuous monitoring of model performance should be implemented to ensure sustained stability and particularly during demand surges or disaster-related scenarios.
3) Although CatBoost shows practical advantages, XGBoost and LightGBM may be retained as

benchmark models for periodic comparative evaluation to ensure continued optimal model selection.

4) Given the operational importance of reducing extreme forecast deviations, the system implementation should prioritize error dispersion monitoring alongside traditional accuracy metrics.
5) The model should be integrated within inventory decision-support mechanisms, particularly for safety stock and reorder point computations, where forecast stability significantly impacts supply chain responsiveness.

Despite the strong predictive performance and robustness demonstrated by CatBoosts, XGBoost, and LightGBM, this study acknowledges an important limitation concerning temporal robustness evaluation. Specifically, a time-based train–test split could not be implemented because the dataset did not contain an explicit time feature. As a result, the evaluation relied on random and repeated K-fold cross-validation techniques rather than chronological data partitioning and it do not fully simulate real-world forecasting conditions where future demand must be predicted strictly from past observations. This limitation suggests that future implementations should incorporate timestamped demand records to enable time-aware validation strategies. Employing chronological validation would further strengthen the assessment of model stability under evolving disaster-related demand environments.

6. Declarations

6.1. Author Contributions

Roman Bariring Villones: Formal analysis, Methodology, Software, Validation, Investigation, Data Curation, Visualization, and Writing – Review & Editing; **Janette Templonuevo Vargas:** Conceptualization, Methodology, Supervision, Project Administration, Writing – Original Draft, and Writing – Review & Editing.

6.2. Institutional Review Board Statement

Ethical review and approval were not required for this study because it did not involve human participants or animals.

6.3. Informed Consent Statement

Informed consent was not required for this study because it did not involve human participants or animals.

6.4. Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

6.5. Acknowledgment

The authors gratefully acknowledge the community in the National Capital Region (NCR), Philippines, for providing access to the datasets used in this study, as well as the Graduate School of La Consolacion University Philippines for its support and guidance throughout the research process.

6.6. Conflicts of Interest

The authors declare no conflicts of interest.

7. References

[1] K. Douaioui, R. Oucheikh, O. Benmoussa, and C. Mabrouki, "Machine Learning and Deep Learning Models for Demand Forecasting in Supply Chain Management: A Critical Review," Applied System Innovation, vol. 7, no. 5, 2024. https://doi.org/10.3390/asi7050093.

[2] I. A. Mohammed and J. Mandal, "Forecasting accuracy through machine learning in supply chain management," *International Journal of Supply Chain Management*, vol. 9, no. 6, pp. 8–24, 2024. https://doi.org/10.47604/ijscm.3074.

[3] S. Pal, "Advancing Multi-Product Warehousing: The Impact and Nuances of Machine Learning-Based Demand Forecasting," *Journal of Supply Chain Analytics*, vol. 7, no. 2, pp. 210–229, 2024.

[4] R. Saha, S. Shofiullah, S. Faysal, and A. Happy, "Systematic Literature Review on Artificial Intelligence Applications In Supply Chain Demand Forecasting," *SSRN Electronic Journal*, paper no. 5062817, 2024. https://dx.doi.org/10.2139/ssrn.5062817.

[5] L. Goel, N. Nandal, S. Gupta, M. Karanam, L. P. Yeluri, A. K. Pandey, O. I. Rozhdestvenskiy, and P. Grabovyg, "Revealing the dynamics of demand forecasting in supply chain management: a holistic investigation," *Cogent Engineering*, vol. 11, no. 1, Art. no. 2368104, 2024. https://doi.org/10.1080/23311916.2024.2368104.

[6] V. P. Parthasarathy, "Machine Learning Algorithms in Supply Chain Forecasting: Accuracy, Efficiency, and Scalability Perspectives," *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, vol. 4, no. 1, pp. 31–39, 2023. https://doi.org/10.63282/3050-9262.IJAIDSML-V4I1P104.

[7] M. A. Jahin, M. S. H. Shovon, J. Shin, I. A. Ridoy, and M. F. Mridha, "Big data-supply chain management framework for forecasting: Data preprocessing and machine learning techniques," *arXiv preprint*, arXiv:2307.12971, 2023. https://doi.org/10.48550/arXiv.2307.12971.

[8] S. Birim, I. Kazancoglu, S. K. Mangla, A. Kahraman, and Y. Kazancoglu, "The derived demand for advertising expenses and implications on sustainability: a comparative study using deep learning and traditional machine learning methods," *Annals of Operations Research*, vol. 339, no. 1, pp. 131–161, 2024. https://doi.org/10.1007/s10479-021-04429-x.

[9] E. M. Onyema, K. K. Almuzaini, F. U. Onu, D. Verma, U. S. Gregory, M. Puttaramaiah, and R. K. Afriyie, "Prospects and challenges of using machine learning for academic forecasting," *Computational Intelligence and Neuroscience*, vol. 2022, Art. no. 5624475, 2022. https://doi.org/10.1155/2022/5624475.

[10] A. ul Husna, S. H. Amin, and A. Ghasempoor, "Machine learning techniques and multi-objective programming to select the best suppliers and determine the orders," *Machine Learning with Applications*, vol. 19, Art. no. 100623, 2025. https://doi.org/10.1016/j.mlwa.2025.100623.

[11] D. Koutsandreas, E. Spiliotis, F. Petropoulos, and V. Assimakopoulos, "On the selection of forecasting accuracy measures," *Journal of the Operational Research Society*, vol. 73, no. 5, pp. 937–954, 2022. https://doi.org/10.1080/01605682.2021.1892464.

[12] C. B. Mupparaju, M. Nalluri, A. S. Rongali, and B. Ananthan, "An Efficient Food Distribution and Logistics Based on Supply Chain," in *Proc. 3rd Int. Conf. Artificial Intelligence for Internet of Things (AIIoT)*, pp. 1–6, IEEE, May 2024. https://doi.org/10.1109/AIIoT58432.2024.10574746.

[13] M. S. Vhatkar, P. S. Mahajan, R. D. Raut, N. Cheikhrouhou, and S. Ghoshal, "Optimized hyperparameters for retail sales forecasting using grid search," *Engineering Applications of Artificial Intelligence*, vol. 158, Art. no. 111472, 2025. https://doi.org/10.1016/j.engappai.2025.111472.

[14] J. Feizabadi, "Machine learning demand forecasting and supply chain performance," *International Journal of Logistics Research and Applications*, vol. 25, no. 2, pp. 119–142, 2022. https://doi.org/10.1080/13675567.2020.1803246.

[15] V. Plotnikova, M. Dumas, and F. P. Milani, "Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements," *Data & Knowledge Engineering*, vol. 139, Art. no. 102013, 2022. https://doi.org/10.1016/j.datak.2022.102013.

[16] A. Mansoori, M. Zeinalnezhad, and L. Nazarimanesh, "Optimization of Tree-Based Machine Learning Models to Predict the Length of Hospital Stay Using Genetic Algorithm," *Journal of Healthcare Engineering*, vol. 2023, Art. no. 9673395, 2023. https://doi.org/10.1155/2023/9673395.

[17] F. Petropoulos et al., "Forecasting: theory and practice," *International Journal of Forecasting*, vol. 38, no. 3, pp. 705–871, 2022. https://doi.org/10.1016/j.ijforecast.2021.11.001.

[18] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021. https://doi.org/10.1016/j.procs.2021.01.199.

[19] M. A. Mediavilla, F. Dietrich, and D. Palm, "Review and analysis of artificial intelligence methods for demand forecasting in supply chain management," *Procedia CIRP*, vol. 107, pp. 1126–1131, 2022. https://doi.org/10.1016/j.procir.2022.05.119.

[20] A. Husna, S. H. Amin, and B. Shah, "Demand forecasting in supply chain management using different deep learning methods," in *Demand Forecasting and Order Planning in Supply*

Chains and Humanitarian Logistics. Hershey, PA, USA: IGI Global, 2021, pp. 140–170. https://doi.org/10.4018/978-1-7998-3805-0.ch005.

[21] A. Forootani, M. Rastegar, and A. Sami, "Short-term individual residential load forecasting using an enhanced machine learning-based approach based on a feature engineering framework: A comparative study with deep learning methods," *Electric Power Systems Research*, vol. 210, Art. no. 108119, 2022. https://doi.org/10.1016/j.epsr.2022.108119.

[22] J. C. Cresswell and T. Kim, "Scaling up diffusion and flow-based XGBoost models," *arXiv preprint*, arXiv:2408.16046, 2024. https://doi.org/10.48550/arXiv.2408.16046.

[23] Z. Zhang, T. Zhang, and J. Li, "Unbiased gradient boosting decision tree with unbiased feature importance," *arXiv preprint*, arXiv:2305.10696, 2023. https://doi.org/10.48550/arXiv.2305.10696.

[24] T. Samal and A. Ghosh, "Ensemble-based predictive analytics for demand forecasting in multi-channel retailing," *Expert Systems with Applications*, Art. no. 130212, 2025. https://doi.org/10.1016/j.eswa.2025.130212.

[25] A. Géron, "Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow," 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2022. https://books.google.co.id/books?id=X5ySEAAAQBAJ.

[26] M. Mohammadagha, "Hyperparameter Optimization Strategies for Tree-Based Machine Learning Models Prediction: A Comparative Study of AdaBoost, Decision Trees, and Random Forest," Decision Trees, and Random Forest, Apr. 11, 2025. https://dx.doi.org/10.2139/ssrn.5226457.

[27] R. G. Goriparthi, "Optimizing Supply Chain Logistics Using AI and Machine Learning Algorithms," *International Journal of Advanced Engineering Technologies and Innovations*, vol. 1, no. 2, pp. 279–298, 2021.

[28] T. O. Hodson, "Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not," *Geosci. Model Dev.*, vol. 2022, pp. 1–10, 2022. https://doi.org/10.5194/gmd-15-5481-2022.

[29] D. S. K. Karunasingha, "Root mean square error or mean absolute error? Use their ratio as well," *Inf. Sci.*, vol. 585, pp. 609–629, 2022. https://doi.org/10.1016/j.ins.2021.11.036.

[30] 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, p. e623, 2021. https://doi.org/10.7717/peerj-cs.623.

[31] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A.-L. Boulesteix, and M. Lindauer, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1484, 2023. https://doi.org/10.1002/widm.1484.

[32] T. Althaqafi, F. Saleem and A. A. M. Al-Ghamdi. "Enhancing student performance prediction: the role of class imbalance handling in machine learning models". *Discover Computing*., vol. 28, no. 1, p. 79, 2025. https://doi.org/10.1007/s10791-025-09576-4.

[33] S. F. Stumpfe and S. C. Shongwe, "Comparative analysis and optimisation of machine learning models for regression and classification on structured tabular datasets," Mathematics, vol. 14, no. 3, p. 473, 2026. https://doi.org/10.3390/math14030473.