International Journal of Computing Sciences Research (ISSN print: 2546-0552; ISSN online: 2546-115X)
Vol. 10, pp. 4100-4118
doi: 10.25147/ijcsr.2017.001.1.261
https://stepacademic.net

Short Paper

Experimental Evaluation of Machine Learning Algorithms for Demand Forecasting of Medical Supplies in Natural Calamity Relief Operations

Roman B. Villones
Graduate School Department, La Consolacion University Philippines, Philippines
roman.villones@email.lcup.edu.ph
(corresponding author)

Jonilo C. Mababa
Graduate School Department, La Consolacion University Philippines, Philippines
jonilo.mababa@email.lcup.edu

Date received: October 20, 2025
Date received in revised form: December 2, 2025; December 5, 2025
Date accepted: December 12, 2025

Recommended citation:

Villones, R. B., & Mababa, J. C. (2026). Experimental evaluation of machine learning algorithms for demand forecasting of medical supplies in natural calamity relief operations. *International Journal of Computing Sciences Research*, *10*, 4100-4118. https://doi.org/ 10.25147/ijcsr.2017.001.1.261

Abstract

Purpose – This study evaluates machine learning algorithms for forecasting medical supply demand in natural calamity relief operations. It investigates forecasting accuracy, robustness, interpretability, and performance.

Method – The Knowledge Discovery in Databases (KDD) process was adopted as the methodological framework. Thirteen algorithms were tested: linear models (Linear Regression, Ridge, Lasso, ElasticNet), tree-based models (Decision Tree, Random Forest, Extra Trees), boosting models (Gradient Boosting, XGBoost, LightGBM, CatBoost), and additional approaches (K-Nearest Neighbors, Support Vector Regressor). Their performance was assessed using RMSE, MAE, and R^2 metrics.

Results – CatBoost achieved the highest baseline R^2 (0.9962), with XGBoost, Extra Trees, LightGBM, and Random Forest also performing strongly (> 0.988). Regression models, KNN, and SVR showed weaker robustness. After hyperparameter tuning with a randomized search and 5-fold cross-validation, LightGBM emerged as the top performer

(0.9941), narrowly surpassing CatBoost and Gradient Boosting, underscoring the advantage of optimized boosting ensembles.

Conclusion – LightGBM, CatBoost, and Gradient Boosting demonstrated superior accuracy and robustness with hyperparameter optimization, further enhancing results.

Recommendations – Disaster response agencies should adopt ensemble models, particularly LightGBM, CatBoost, and Gradient Boosting, within their decision-support systems, while applying hyperparameter tuning and exploring real-time data integration for future applications.

Research Implications – Findings reinforce boosting-based ensembles as reliable tools for disaster demand forecasting. Enhancing academic understanding and improving logistics by efficiency, response times, and resource allocation in relief operations.

Keywords – Machine Learning, Demand Forecasting, Medical Supplies, Disaster Relief Operations, Hyperparameter Optimization

INTRODUCTION

Natural calamities such as typhoons, earthquakes, floods, and volcanic eruptions frequently disrupt communities, leading to massive displacement and health emergencies. In such situations, the essential medical supplies, including protective equipment and first-aid kits, are playing a vital role in saving lives and preventing further health crises. The unpredictability of disasters often causes sudden spikes in demand, creating logistical challenges for relief agencies. Without accurate forecasting, shortages or oversupply may occur, which can delay critical interventions and reduce the effectiveness of relief operations, according to Jebbor et al. (2022). Ensuring the timely availability and distribution of medical supplies is therefore a cornerstone of effective disaster response.

According to Yani and Aamer (2023), machine learning (ML) offers innovative solutions to these challenges by enabling data-driven forecasting of demand under uncertain and dynamic conditions. Unlike traditional forecasting methods, ML can capture complex and non-linear relationships in data, making it highly suitable for volatile environments such as disaster relief. In the study of Lin et al. (2025), advanced algorithms such as ensemble methods and tree-based algorithms have demonstrated superior accuracy in predicting healthcare supply needs. The integration of ML into disaster logistics not only improves demand forecasting but also enhances decision-making, optimizing resource allocation, and reducing wastage during relief efforts.

This study aims to experimentally evaluate multiple machine learning algorithms for forecasting the demand of medical supplies in natural calamity relief operations. Specifically, the study seeks to determine which algorithms provide the highest forecasting accuracy, assess their robustness under disaster-related data challenges, and

identify practical trade-offs between interpretability and performance. By providing a comparative analysis, the study intends to guide disaster response agencies in selecting the most effective algorithmic approaches for real-world relief operations.

This study offers both academic and practical contributions to the field of disaster logistics and data-driven decision support. Academically, it advances the literature by providing a rigorous empirical comparison of multiple machine learning algorithms for forecasting medical supply demand in the unique context of natural calamity relief operations. Practically, the findings equip disaster response agencies and humanitarian organizations with evidence-based guidance on selecting forecasting models that balance accuracy, interpretability, and operational feasibility. By identifying which algorithms perform most reliably during crises and highlighting the trade-offs associated with each. The study supports more informed decision-making, improved pre-positioning of medical supplies, and greater efficiency in real-world relief operations.

LITERATURE REVIEW

Machine learning has been increasingly applied to the forecasting of medical supply demand, particularly in contexts where uncertainty and rapid changes in consumption patterns occur. For instance, the study of Mbonyinshuti et al. (2021) in Rwanda used Random Forest to forecast essential medicine demand with notable accuracy and demonstrated its ability to capture complex patterns in consumption data. Similarly, reported that Random Forest and decision tree algorithms achieved up to 41% higher accuracy than traditional approaches in pharmaceutical supply chain forecasting, as per Yani and Aamer (2023). In more specialized contexts, Lin et al. (2025) showed that XGBoost could reliably forecast emergency drug demand following earthquakes by underscoring the robustness of ensemble and boosting techniques in volatile environments. While these studies highlight the strength of advanced algorithms, even regularized regressions like Ridge, Lasso, and ElasticNet remain competitive, especially when datasets are large and rich in predictive features, as explained by Vollmer et al. (2021).

While algorithm choice is crucial, the reliability of predictions depends significantly on the validation strategy. Wilimitis and Walsh (2023) examined different cross-validation strategies in healthcare machine learning applications. They emphasized that repeated K-Fold cross-validation provides more stable and less biased performance estimates compared to simple hold-out or single K-Fold. Similarly, Maldonado et al. (2022) demonstrated that repeated cross-validation provides more consistent performance evaluation compared to single holdout methods, especially in clinical time-series prediction tasks, thereby improving the robustness of healthcare forecasting models.

Another factor central to model accuracy is hyperparameter optimization. Among available techniques, Hidayaturrohman and Hanada (2025) found that Random Search achieved performance comparable to Grid Search, while Bayesian Search required less

computation time when tuning hyperparameters for Random Forest, XGBoost, and SVM models in predicting heart failure outcomes. More recently, Meaney et al. (2025) demonstrated that RandomizedSearchCV significantly improved both discrimination and calibration in XGBoost models used to predict high-need and high-cost healthcare users. Similarly, Lamir et al. (2025) reported that RandomizedSearchCV tuned Random Forest models achieved higher predictive accuracy for heart disease forecasting than grid-tuned models, while reducing computational demands. Pathak et al. (2024) advanced this line of research by introducing a hybrid "Randomized-Grid Search" method, showing that randomization improves search efficiency without sacrificing accuracy.

To improve the models' ability to capture complex dynamics in medical supply distribution during disaster relief operations. According to Monsef et al. (2023), applying supply–demand forecast modelling to identify healthcare service gaps in Dubai and effectively quantifying shortages and surpluses over time to guide resource allocation and policy planning. As explained by Praneetpholkrang and Kanjanawattana (2021), the Weighted Demand Index was added to reflect the prioritization of certain medical supplies during calamities since not all items contribute equally to relief operations. Cabello-Solorzano et al. (2023, August) emphasized that normalization techniques, such as min–max scaling and z-score standardization, significantly improve the stability and accuracy of machine learning models like KNN and SVM. Finally, Wang (2022) introduced a composite demand indicator that merges multiple demand signals (e.g., historical sales, web search interest, promotional activities) into a single index, which improves forecast interpretability and enhances robustness against fluctuations in individual indicators.

Recent literature shows that machine learning has become a critical tool in forecasting medical supply demand, especially in environments where consumption patterns shift rapidly during disaster-driven scenarios. Ensemble methods such as Random Forest and XGBoost consistently demonstrate superior predictive performance compared to traditional forecasting approaches. Evidence from studies conducted in Rwanda and across the broader pharmaceutical supply chain reports accuracy gains as high as 41%. However, regularized regression models like Ridge, Lasso, and ElasticNet remain valuable when datasets are large and rich in predictive features, demonstrating that algorithm effectiveness depends heavily on the structure and characteristics of the available data. These models provide stable performance in high-dimensional environments and offer interpretability advantages not always present in more complex ensembles.

Beyond algorithm selection, recent studies emphasize the importance of rigorous validation strategies. Repeated K-Fold cross-validation, as shown by Wilimitis and Walsh (2023) and Maldonado et al. (2022), provides a more stable and unbiased estimate of model performance compared to simple hold-out or single K-Fold methods. This ensures that predictive accuracy is measured more reliably across varying data partitions. Hyperparameter optimization further enhances model performance. Studies by Hidayaturrohman and Hanada (2025), Meaney et al. (2025), and Lamir et al. (2025)

demonstrate that RandomizedSearchCV and Bayesian optimization approaches reduce computational costs while maintaining or improving predictive accuracy compared to exhaustive grid search. Hybrid techniques, such as the randomized-grid search method proposed by Pathak et al. (2024), also show that introducing randomization increases efficiency without sacrificing accuracy.

Finally, research highlights the need for forecasting models that account for the unique operational complexities of medical supply distribution during disasters. Frameworks such as the supply–demand gap model of Monsef et al. (2023) and the Weighted Demand Index developed by Praneetpholkrang and Kanjanawattana (2021) offer mechanisms for prioritizing critical medical items during relief operations. Complementary methods, including normalization techniques by Cabello-Solorzano et al. (2023) and composite demand indicators proposed by Wang (2022), further enhance model stability and interpretability in the presence of fluctuating demand patterns. Together, these studies indicate that accurate medical supply forecasting depends not only on advanced algorithms but also on robust validation, efficient optimization, and thoughtful feature engineering tailored to dynamic healthcare environments.

METHODOLOGY

Knowledge Discoveries in Databases

This study adopted the Knowledge Discovery in Databases (KDD) process as the methodological framework for the experimental evaluation of machine learning algorithms in forecasting the demand for medical supplies during natural calamity relief operations. According to Mavrogiorgou et al. (2021, May), the KDD process provided a systematic approach for handling data from acquisition to knowledge generation and ensuring both rigor and reproducibility.

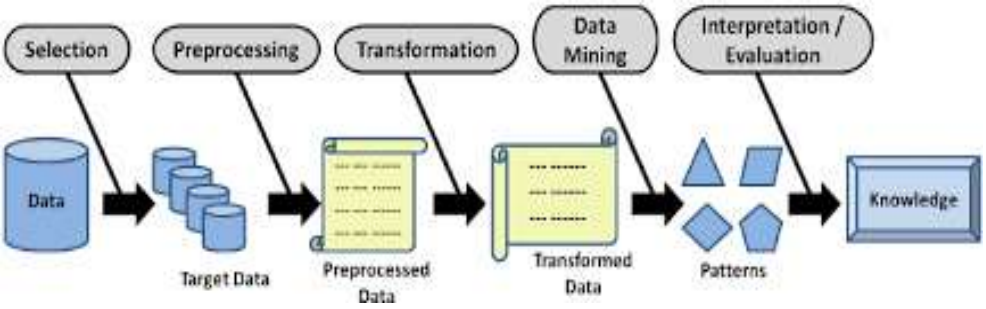

Figure 1. Gullo, F. (2015). KDD Process (as cited by Umam & Handayani, 2025).

Figure 1 illustrates the Knowledge Discovery in Databases (KDD) process, which outlines the systematic steps followed in this study. It begins with the selection and preprocessing of disaster-related data, then continues with transformation to prepare the dataset for analysis, and proceeds to data mining to uncover meaningful patterns. The final stage involves interpretation and evaluation, where the extracted knowledge is

translated into actionable insights to support decision-making in medical supply demand forecasting during natural calamities.

Selection

The dataset utilized in this study was obtained from selected communities within the National Capital Region (NCR), Philippines, using publicly available sources related to medical supply distribution during natural calamities. The dataset covers the period from January 2018 to December 2023 and captures multiple disaster response operations in the region. It consists of more than 76,000 records with ten attributes and is categorized into both categorical and numerical variables.

The categorical variables include Item ID, which identifies various medicines and medical equipment such as antibiotics, analgesics, antipyretics, rehydration solutions, personal protective equipment, first-aid instruments, and basic diagnostic tools. Additional categorical fields include District, indicating the specific geographical area within NCR, and the type of Calamities by describing the type of disaster encountered, such as typhoons, floods, earthquakes, and volcanic eruptions.

The numerical variables consist of Inventory Level, representing the available quantity of each medical item at the time of recording. Relief Aid refers to the total quantity of supplies distributed during disaster-response operations. And Inventory Re-Ordered, indicating the volume of stock replenished based on demand triggers. Additional derived measures include the Simple Ratio Method, a proportional value relating available inventory to historical usage. The Weighted Demand Index, which integrates past consumption patterns and calamity frequency to estimate adjusted demand intensity. The Normalized Score is a standardized metric that rescales demand-related values for comparability. The Overall Demand Score is a composite indicator summarizing inventory behavior, distribution patterns, and demand metrics to determine each item's relative priority during calamity operations.

Pre-processing

The dataset did not contain any missing values, which reduced the need for using other techniques and allowed direct progression to feature enhancement. Figure 2 presents the distribution of inventory-related variables. The Inventory Level shows a right-skewed distribution where most items are concentrated at lower levels with a long tail extending to higher quantities, indicating that high stock levels are rare. The Relief Aid distribution appears moderately skewed, with most values clustering around 50–100 units and suggesting a common aid allocation range, while extreme values remain uncommon. Meanwhile, the Inventory Re-Ordered distribution is highly skewed, with the majority of re-order quantities being very small and only a few instances of large re-orders.

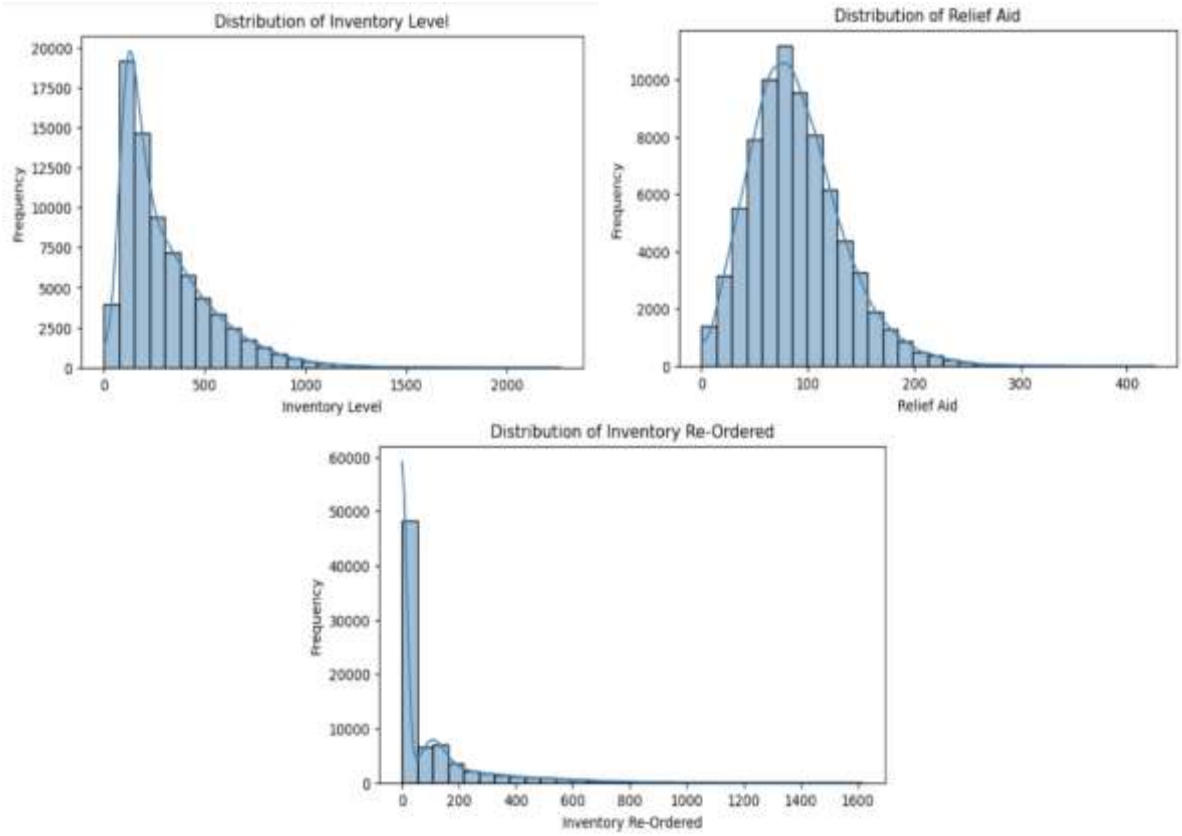

Figure 2. Distribution of Inventory Level, Relief Aid, and Inventory Re-Ordered

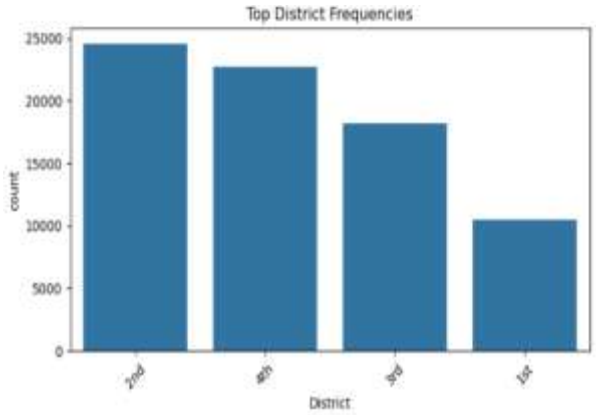

Figure 3. Top District Frequencies

Figure 3 illustrates the frequency distribution of districts. It shows that the 2nd District recorded the highest count, followed closely by the 4th and 3rd Districts, while the 1st District had the lowest representation. This suggests that relief operations and inventory activities are more concentrated in certain districts, possibly due to population density or exposure to risks.

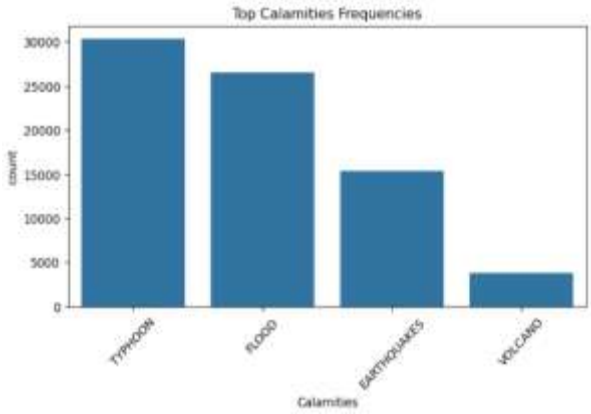

Figure 4. Top Calamities Frequencies

Figure 4 illustrates the frequency distribution of calamities. It highlights that typhoons and floods are the most frequent calamities encountered, with earthquakes occurring less often and volcanic events being relatively rare. These results emphasize that disaster response efforts must prioritize weather-related calamities while still maintaining preparedness for less frequent but potentially severe hazards.

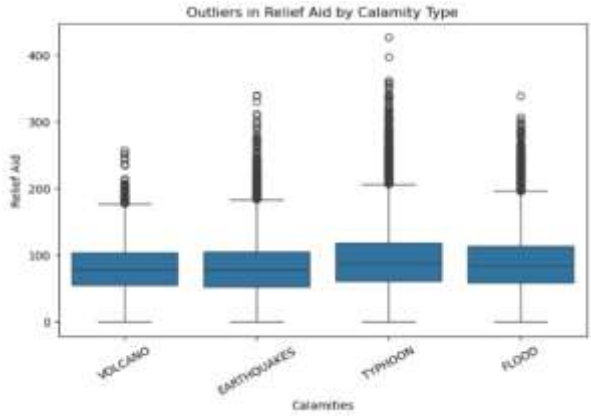

Figure 5. Outliers in Relief Aid by Calamity Type

Figure 5 displays the distribution of relief aid across different calamity types using a box plot. The results show that the median relief aid provided is relatively consistent across calamities and falls within the range of 70–100 units. However, the spread of values and the presence of outliers vary significantly. Typhoons and floods exhibit a wider distribution with numerous high-value outliers, reflecting occasional large-scale relief operations. Earthquakes also display a moderate spread with several extreme values, while volcanic events tend to have lower variability and fewer large outliers.

Outliers were examined using the interquartile range (IQR) method to determine whether they represented data errors. Since these extreme values corresponded to legitimate large-scale relief deployments during severe calamities, they were retained in the dataset rather than removed. However, they were treated carefully during analysis by

applying normalization techniques and robust statistical measures to prevent them from disproportionately influencing the model performance and interpretation.

Transformation

In this phase, the categorical attributes were converted into numerical form using Label Encoding, which assigns unique integer values to each category. This transformation ensured that categorical information could be effectively utilized by machine learning algorithms. The target variable for prediction was defined as Relief Aid, representing the demand for medical supplies during natural disasters. The remaining attributes were treated as features to support forecasting.

To prepare the dataset for modeling, the records were split into training and testing subsets using an 80:20 ratio. The training set was used to build and optimize the models, while the testing set was reserved for independent performance evaluation. This partitioning allowed for a fair assessment of the algorithm's ability to generalize to the unseen data.

The study utilized several evaluation metrics to assess the predictive performance and reliability of the machine learning algorithms. Mean Absolute Error (MAE) was employed to measure the average magnitude of prediction errors by providing an intuitive indication of how closely the model's estimates aligned with actual demand values. Root Mean Squared Error (RMSE) quantifies the square root of the average squared differences between predicted and observed values by placing greater emphasis on larger errors and offering insight into the model's sensitivity to substantial deviations. The Coefficient of Determination (R^2) was applied to evaluate the proportion of variance in the dependent variable explained by the model and serves as a key indicator of model fit and explanatory power.

To ensure robust and unbiased performance assessment, the study implemented 5-Fold Cross-Validation, in which the dataset was partitioned into five subsets, allowing the model to be iteratively trained and tested across multiple splits. The resulting R^2 Mean and R^2 Standard Deviation were computed to summarize the model's average predictive accuracy and the consistency of its performance across all folds.

Randomized Search Cross-Validation, or RandomizedSearchCV, is a hyperparameter optimization technique used to identify the most effective combination of hyperparameters for a machine learning algorithm. Unlike exhaustive grid search, which evaluates all possible parameter combinations, the RandomizedSearchCV samples a fixed number of parameter settings from specified distributions. This method is often coupled with k-fold cross-validation, which splits the dataset into multiple folds to assess model performance across different subsets. RandomizedSearchCV efficiency and flexibility make it particularly suitable for large datasets and complex models, and it has been used in forecasting medical supply demand in disaster relief operations.

Data Mining

In this phase, a diverse set of thirteen machine learning algorithms was implemented to forecast medical supply demand during natural disasters. The algorithms were carefully selected to represent different families of learning approaches and ensure a comprehensive experimental comparison.

Linear algorithms, including Linear Regression, Ridge Regression, Lasso Regression, and ElasticNet, were employed as baseline learners. These models offered interpretable regression outputs by enabling a clear understanding of how input variables influenced predicted demand. Their inclusion also allowed for the evaluation of regularization techniques and their effectiveness in controlling overfitting, reducing model complexity, and improving generalization performance.

Followed by, Tree-based algorithms such as Decision Tree, Random Forest, and Extra Trees were selected for their capacity to capture non-linear relationships and handle complex feature interactions. These models are inherently robust, particularly when dealing with diverse or noisy datasets common in disaster-related scenarios. The ensemble variants further enhanced predictive stability by reducing variance and increasing model reliability compared to a single decision tree.

Furthermore, boosting algorithms like Gradient Boosting, XGBoost, LightGBM, and CatBoost have formed a central component of the experimental framework due to their demonstrated effectiveness in structured tabular datasets. These models iteratively refine prediction errors from previous weak learners, resulting in strong predictive accuracy and computational efficiency. Their performance characteristics make them well-suited for real-world forecasting applications, especially in time-sensitive and high-stakes environments such as disaster relief operations.

Finally, two additional algorithms, K-Nearest Neighbors (KNN) and Support Vector Regressor (SVR), were incorporated to provide alternative modeling perspectives. KNN was selected for its simplicity and non-parametric nature by offering predictions based on similarity patterns within the dataset. In contrast, SVR was included for its ability to handle high-dimensional feature spaces and capture complex nonlinear relationships through kernel-based transformations. Together, these models broadened the analytical landscape and enabled a more comprehensive evaluation across diverse learning paradigms.

These algorithms ensured a balanced evaluation of various modeling paradigms by supporting the objective of identifying the most suitable approach for reliable and accurate demand forecasting of medical supplies during calamity response operations.

RESULTS

Evaluation

The performance of the machine learning algorithms was assessed using multiple evaluation metrics to ensure a comprehensive understanding of predictive accuracy. Specifically, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) were employed.

Table 1. Performance Comparison of Machine Learning Algorithms.

#	Algorithms	MAE	RSME	R^2	Rank
11	CatBoost	1.630896	2.721914	0.996167	1
9	XGBoost	2.482483	4.211169	0.990826	2
7	Extra Trees	1.578415	4.248396	0.990663	3
10	LightGBM	2.749605	4.673677	0.988700	4
6	Random Forest	1.805452	4.843289	0.987865	5
5	Decision Tree	3.016316	7.418307	0.971531	6
8	Gradient Boosting	6.753341	11.998922	0.925519	7
1	Linear Regression	22.192781	30.113811	0.530872	8
2	Ridge Regression	22.200398	30.114210	0.530859	9
13	Support Vector Regressor	27.760314	36.123928	0.324928	10
12	K-Nearest Neighbors	26.997461	36.452364	0.312597	11
3	Lasso Regression	28.859528	36.757216	0.301051	12
4	ElasticNet	29.454988	37.440241	0.274834	13

Table 1 presents the evaluation results, which indicate that ensemble and boosting-based algorithms substantially outperformed traditional regression models. CatBoost achieved the best performance, with the lowest MAE (1.63) and RMSE (2.72) alongside the highest R^2 (0.996) and demonstrating strong predictive accuracy and generalization. XGBoost, Extra Trees, LightGBM, and Random Forest followed closely, and all maintained R^2 scores above 0.987, underscoring the robustness of ensemble methods in capturing complex non-linear relationships. In contrast, single-tree models such as Decision Tree and Gradient Boosting are performed moderately, while traditional linear models like Linear Regression, Ridge, Lasso, and ElasticNet, and kernel-based approaches KNN and SVR produced high error values and low explanatory power by confirming their limited suitability for this dataset.

To reduce bias and variance in model evaluation, K-Fold Cross-Validation was applied. In this approach, the dataset was partitioned into k equally sized subsets, with each fold used once as a validation set while the remaining folds were used for training.

Table 2 shows that CatBoost consistently achieved the highest predictive accuracy with a mean R^2 (0.9962) and minimal variance (0.0003), indicating strong generalization. XGBoost, Extra Trees, LightGBM, and Random Forest also performed reliably with mean R^2 values above (0.988 and relatively low standard deviations. Decision Tree and Gradient Boosting demonstrated moderate performance, while Linear Regression, Ridge Regression, Lasso Regression, ElasticNet, K-Nearest Neighbors, and Support Vector

Regressor produced lower mean R^2 scores and higher variability, underscoring their limited robustness for demand forecasting in this context.

Table 2. K-Fold (k=5) Cross-Validation Results for Machine Learning Algorithms.

#	Algorithms	R^2 Mean	R^2 std	Rank
11	CatBoost	0.996180	0.000277	1
9	XGBoost	0.991664	0.000636	2
7	Extra Trees	0.990501	0.001932	3
10	LightGBM	0.988892	0.000909	4
6	Random Forest	0.988377	0.001345	5
5	Decision Tree	0.970971	0.001654	6
8	Gradient Boosting	0.923976	0.004348	7
1	Linear Regression	0.521444	0.013612	8
2	Ridge Regression	0.521427	0.013693	9
13	Support Vector Regressor	0.302919	0.008395	10
12	K-Nearest Neighbors	0.293703	0.016683	11
3	Lasso Regression	0.284519	0.010821	12
4	ElasticNet	0.257347	0.010304	13

For hyperparameter optimization, RandomizedSearchCV was employed to efficiently explore the parameter space of the top-performing algorithms. And only the algorithms that demonstrated strong baseline performance, such as CatBoost, XGBoost, Extra Trees, LightGBM, Random Forest, Decision Tree, and Gradient Boosting, were included in this optimization process.

Table 3. Model Performance Before and After Hyperparameter Optimization.

#	Algorithms	R^2 (Baseline)	R^2 (Tuned)	Difference	Rank
7	CatBoost	0.996167	0.994223	-0.001944	1
6	LightGBM	0.988700	0.993827	0.005127	2
4	Gradient Boosting	0.925519	0.993382	0.067863	3
5	XGBoost	0.990826	0.991869	0.001043	4
3	Extra Trees	0.990663	0.990438	-0.000225	5
2	Random Forest	0.987865	0.987736	-0.000129	6
1	Decision Tree	0.971531	0.971900	0.000369	7

Table 3 presents the top-performing algorithms from the earlier evaluations, which were further subjected to hyperparameter tuning. The optimization process had varying effects across models. CatBoost maintained its leading position; however, its tuned performance showed a slight decrease in R^2 (from 0.9962 to 0.9942). In contrast, LightGBM improved in R^2 (from 0.9887 to 0.9938), and Gradient Boosting showed the most significant gain in R^2 (from 0.9255 to 0.9934). XGBoost also recorded a modest

improvement by increasing R^2 (from 0.9908 to 0.9919). Extra Trees showed a slight decrease in R^2 (from 0.9907 to 0.9904), and Random Forest remained nearly unchanged by declining marginally in R^2 (from 0.9879 to 0.9877). Decision Tree exhibited a minor increase in R^2 (from 0.9715 to 0.9719).

These results indicate that while hyperparameter tuning can substantially improve algorithm performance, the sensitivity to parameter adjustments differs depending on the algorithm's characteristics. Models such as CatBoost, which rely on highly optimized default settings and built-in regularization, require a carefully designed search space to achieve further gains without risking slight declines. LightGBM and Gradient Boosting are more responsive to tuning by demonstrating that the models can benefit substantially from parameter optimization. XGBoost shows limited but positive improvements, suggesting that its defaults are already strong but can gain from fine-tuning. Ensemble bagging models like Extra Trees and Random Forest exhibit minimal sensitivity to tuning by reflecting their inherent stability and robustness. Decision Trees demonstrate minor responsiveness, as their simpler structure limits the impact of parameter optimization compared to ensemble methods.

Table 4. K-Fold (k=5) Cross-Validation Results after applying Best Hyperparameters for each algorithm.

#	Algorithms	R^2 Mean	R^2 std	Rank
6	LightGBM	0.994083	0.000828	1
7	CatBoost	0.993896	0.000263	2
4	Gradient Boosting	0.993598	0.000642	3
5	XGBoost	0.992253	0.000447	4
3	Extra Trees	0.990142	0.001625	5
2	Random Forest	0.987929	0.001147	6
1	Decision Tree	0.972859	0.001824	7

Table 4 shows the following hyperparameter tuning with a 5-fold cross-validation. As presented, the results show that LightGBM achieved the highest mean R^2 of 0.9941 with a standard deviation of 0.0008. Followed closely by CatBoost with a mean R^2 of 0.9939 and a very low standard deviation of 0.0003. Gradient Boosting attained a mean R^2 of 0.9936 with a standard deviation of 0.0006. While XGBoost achieved a mean R^2 of 0.9923 and a standard deviation of 0.0004. Extra Trees achieved a mean R^2 value of 0.9901 and a standard deviation of 0.0016. Random Forest recorded slightly lower mean R^2 values of 0.9879 and a standard deviation of 0.0011. Finally, the Decision Tree showed the lowest mean R^2 of 0.9729 with a standard deviation of 0.0018.

These results indicate that while all tuned algorithms performed well, their sensitivity to parameter optimization and stability across folds varied. Gradient boosting-based models, such as LightGBM, CatBoost, and Gradient Boosting, not only achieved the highest predictive accuracy but also demonstrated a low variability, highlighting their robustness and responsiveness to hyperparameter tuning. XGBoost showed moderately

high performance with limited variation and suggests an effective default with additional benefit from tuning. Bagging-based models like Extra Trees and Random Forest displayed strong performance but slightly higher variability, reflecting their stability but limited sensitivity to tuning. Decision Tree exhibited the lowest accuracy and the highest variation by underscoring its susceptibility to overfitting and the need for ensemble approaches to achieve robust predictions.

Table 5. Best Parameters for each algorithm.

Rank	Algorithms	Best Parameters
1	LightGBM	'num_leaves': 127, 'n_estimators': 200, 'learning_rate': 0.05
2	CatBoost	'learning_rate': 0.2, 'iterations': 500, 'depth': 4
3	Gradient Boosting	'n_estimators': 200, 'max_depth': 7, 'learning_rate': 0.05
4	XGBoost	'n_estimators': 200, 'max_depth': 7, 'learning_rate': 0.05
5	Extra Trees	'n_estimators': 100, 'min_samples_split': 5, 'max_depth': 30
6	Random Forest	'n_estimators': 100, 'min_samples_split': 5, 'max_depth': 30
7	Decision Tree	'min_samples_split': 2, 'min_samples_leaf': 2, 'max_depth': 20

Table 5 shows that the hyperparameter optimization process identified a distinct set of best-performing configurations for each algorithm. For the boosting-based models, LightGBM achieved optimal performance with a relatively high number of leaves of 127 and 200 estimators combined with a conservative learning rate of 0.05, indicating that deeper leaf-wise growth paired with gradual learning improved its ability to capture complex demand patterns. On the other hand, CatBoost performed its best parameters with a higher learning rate of 0.2, 500 iterations, and a shallow depth of 4, suggesting that frequent but shallow updates were more effective for its categorical feature handling and gradient boosting structure. Gradient Boosting and XGBoost converged on a similar parameter set of 200 estimators, a max depth of 7, and a learning rate of 0.05, highlighting that moderate depth and stable incremental learning supported strong generalization across these tree-based boosting frameworks. For ensemble bagging models, Extra Trees and Random Forest shared optimal parameters, with 100 estimators, a minimum sample split of 5, and a deeper maximum depth of 30. This configuration reflects their reliance on diverse deep trees to capture variance within the dataset. Lastly, the Decision Tree model achieved its best performance with a max depth of 20, and minimum sample thresholds of 2 for splitting and leaf formation, indicating that a relatively deep but minimally constrained tree structure was most effective in modeling the underlying relationships without excessive overfitting. Together, these optimized configurations illustrate how different algorithms require tailored hyperparameter settings to maximize their predictive accuracy in forecasting medical supply demand.

DISCUSSION

The findings clearly indicate that ensemble and boosting-based algorithms consistently outperformed traditional regression and distance-based approaches. CatBoost demonstrated strong predictive accuracy and stability, while other ensemble methods such as XGBoost, Extra Trees, LightGBM, and Random Forest also produced reliable results. These models proved effective in handling the non-linear and complex nature of disaster relief data and making them highly suitable for demand forecasting. In contrast, linear regression variants and kernel-based methods were less effective, as they struggled with the variability and complexity inherent in the dataset.

The application of hyperparameter optimization through RandomizedSearchCV further highlighted the adaptability of boosting-based methods. The LightGBM benefited from tuning and slightly surpassed CatBoost, while Gradient Boosting showed marked improvement compared to its baseline performance. XGBoost recorded the modest gains, whereas Extra Trees and Random Forest exhibited limited changes. CatBoost's slight performance decrease after tuning underscores the importance of a carefully designed search space for already highly optimized models. These outcomes reinforce that hyperparameter tuning can significantly enhance model reliability and fully leverage the predictive power of machine learning algorithms.

Feature importance analysis was performed for all evaluated algorithms. For LightGBM, CatBoost, and Gradient Boosting, the analysis revealed that Inventory Level, Relief Aid, and District were the most influential features, followed by Calamity Type and derived metrics such as the Weighted Demand Index and Overall Demand Score. This allows the extraction of relative feature importance and provides insight into which variables most strongly drive demand forecasts. For XGBoost, Extra Trees, and Random Forest, similar patterns were observed. The algorithms consistently highlighted the Inventory Level and Relief Aid as the primary determinants, and District and Calamity Type also contributed substantially. The Decision Tree provided a more straightforward and transparent hierarchy of feature splits, confirming the relative importance of the same key variables. Across all algorithms, the findings demonstrate that ensemble methods can generate clear and actionable insights into the key factors influencing medical supply demand and effectively support data-driven forecasting and decision-making in disaster response operations.

Overall, the study demonstrates that boosting-based ensemble methods, particularly LightGBM, CatBoost, and Gradient Boosting, are the most appropriate for forecasting medical supply demand in disaster scenarios. Their strength lies in balancing accuracy and robustness by ensuring more reliable predictions that can guide decision-makers in allocating resources effectively. In operational contexts where timely and accurate forecasting is critical, the more complex models offer greater practical value despite their reduced transparency.

CONCLUSIONS AND RECOMMENDATIONS

In conclusion, the study successfully achieved its objectives by experimentally evaluating multiple machine learning algorithms for forecasting medical supply demand in natural calamity relief operations. The experimentation confirmed that LightGBM, CatBoost, and Gradient Boosting consistently delivered the highest forecasting accuracy by identifying top-performing algorithms. The evaluation under disaster-related data challenges demonstrated that these models maintained strong predictive performance, indicating that the objective of assessing robustness was effectively met. Furthermore, the comparative analysis provided clarity on the practical trade-offs between interpretability and performance with boosting methods. It offers both strong accuracy and meaningful insights into influential demand factors. Through these findings, the study fulfills its aim of guiding disaster response agencies toward selecting effective and reliable algorithmic approaches for real-world relief operations.

It is recommended that disaster response agencies and humanitarian organizations adopt advanced ensemble models such as LightGBM, CatBoost, and Gradient Boosting for forecasting medical supply demand during natural disasters. These algorithms should be integrated into decision-support systems to enhance the accuracy and timeliness of resource allocation, thereby reducing the risk of shortages in critical relief operations. Furthermore, the use of hyperparameter optimization techniques, such as RandomizedSearchCV, is advised to maximize model performance and adaptability across varying disaster scenarios. For future implementations, researchers may also explore combining these models with real-time data streams to further improve predictive reliability and operational responsiveness.

IMPLICATIONS

The results of this study carry important implications for both research and practice. For the academic community, the findings reinforce the potential of boosting-based ensemble algorithms as reliable approaches for demand forecasting in complex and high-variability domains like disaster management. From a practical standpoint, the study provides evidence that disaster response agencies can strengthen relief operations by integrating machine learning into their logistics systems. More accurate forecasts of medical supply needs can lead to faster response times, reduced shortages, and more efficient allocation of limited resources during natural disasters.

ACKNOWLEDGEMENT

The researchers gratefully acknowledge the community in the NCR, Philippines, who provided access to their datasets, which played a vital role in this study. And the Graduate School of La Consolacion University Philippines for its support and guidance throughout the research.

FUNDING

This research did not receive any specific grant from any funding agencies.

DECLARATIONS

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this study.

Informed Consent

The datasets used in this study were obtained from the community in the NCR, Philippines; their terms of use permit uploaded data to be utilized for research purposes; therefore, informed consent was not applicable.

Ethics Approval

As the data were obtained from the community in the NCR, Philippines, ethics approval was not required, since the dataset is intended for research use and does not contain any sensitive or personally identifiable information.

REFERENCES

Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., & J. Tallón-Ballesteros, A. (2023, August). The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis. In International Conference on Soft Computing Models in Industrial and Environmental Applications (pp. 344-353). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42536-3_33

Gullo, F. (2015). From patterns in data to knowledge discovery: What data mining can do. *Physics Procedia, 62*, 18-22. https://doi.org/10.1016/j.phpro.2015.02.005

Hidayaturrohman, Q. A., & Hanada, E. (2025). A Comparative Analysis of Hyper-Parameter Optimization Methods for Predicting Heart Failure Outcomes. Applied Sciences, 15(6), 3393. https://doi.org/10.3390/app15063393

Jebbor, S., Raddouane, C., & El Afia, A. (2022). A preliminary study for selecting the appropriate AI-based forecasting model for hospital assets demand under disasters. Journal of Humanitarian Logistics and Supply Chain Management, 12(1), 1-29. https://doi.org/10.1108/JHLSCM-12-2020-0123

Lamir, A. A., Razzagzadeh, S., & Rezaei, Z. (2025). A Comprehensive Machine Learning Framework for Heart Disease Prediction: Performance Evaluation and Future Perspectives. https://doi.org/10.48550/arXiv.2505.09969

Lin, Y., Yan, W., Zhang, Y., Zhang, H., Zhang, H., & Wang, D. (2025). Emergency Drug Demand Forecasting in Earthquakes with XGBoost and AFT-LSTM. Sustainability, 17(5), 1910. https://doi.org/10.3390/su17051910

Maldonado, S., López, J., & Iturriaga, A. (2022). Out-of-time cross-validation strategies for classification in the presence of dataset shift. Applied Intelligence, 52(5), 5770-5783. https://doi.org/10.1007/s10489-021-02735-2

Mavrogiorgou, A., Kiourtis, A., Manias, G., & Kyriazis, D. (2021, May). An optimized KDD process for collecting and processing ingested and streaming healthcare data. In 2021, the 12th international conference on information and communication systems (ICICS) (pp. 49-56). IEEE. https://doi.org/10.1109/ICICS52457.2021.9464551

Mbonyinshuti, F., Nkurunziza, J., Niyobuhungiro, J., & Kayitare, E. (2021). The prediction of essential medicines demand: a machine learning approach using consumption data in Rwanda. Processes, 10(1), 26. https://doi.org/10.3390/pr10010026

Meaney, C., Wang, X., Guan, J., & Stukel, T. A. (2025). Comparison of methods for tuning machine learning model hyperparameters: with application to predicting high-need high-cost health care users. BMC Medical Research Methodology, 25(1), 134. https://doi.org/10.1186/s12874-025-02561-x

Monsef, N., Suliman, E., Ashkar, E., & Hussain, H. Y. (2023). Healthcare services gap analysis: a supply capture and demand forecast modelling, Dubai 2018–2030. BMC Health Services Research, 23(1), 468. https://doi.org/10.1186/s12913-023-09401-y

Pathak, A. K., Chaubey, M., & Gupta, M. (2024). Randomized-Grid Search for Hyperparameter Tuning in Decision Tree Model to Improve Performance of Cardiovascular Disease Classification. https://doi.org/10.48550/arXiv.2411.18234

Praneetpholkrang, P., & Kanjanawattana, S. (2021). A multi-objective optimization model for shelter location-allocation in response to humanitarian relief logistics. The Asian Journal of Shipping and Logistics, 37(2), 149-156. https://doi.org/10.1016/j.ajsl.2021.01.003

Umam, K., & Handayani, M. R. (2025). Performance of Machine Learning Algorithms on Imbalanced Sentiment Datasets Without Balancing Techniques. Journal of Applied Informatics and Computing, 9(3), 998-1005. https://doi.org/10.30871/jaic.v9i3.9584

Vollmer, M. A., Glampson, B., Mellan, T., Mishra, S., Mercuri, L., Costello, C., Klaber, R., Cooke, G., Flaxman, S., & Bhatt, S. (2021). A unified machine learning approach to time series forecasting applied to demand at emergency departments. BMC Emergency Medicine, 21(1), 9. https://doi.org/10.1186/s12873-020-00395-y

Wang, C. H. (2022). Considering economic indicators and dynamic channel interactions to conduct sales forecasting for retail sectors. Computers & Industrial Engineering, 165, 107965. https://doi.org/10.1016/j.cie.2022.107965

Wilimitis, D., & Walsh, C. G. (2023). Practical considerations and applied examples of cross-validation for model development and evaluation in health care: tutorial. Jmir ai, 2, e49023. https://doi.org/10.2196/49023

Yani, L. P. E., & Aamer, A. (2023). Demand forecasting accuracy in the pharmaceutical supply chain: a machine learning approach. International journal of pharmaceutical and healthcare marketing, 17(1), 1-23. https://doi.org/10.1108/IJPHM-05-2021-0056

Author's Biography

Roman B. Villones is an Assistant Professor at the College of Informatics, Philippine Christian University. He holds a Master's degree in Information Technology and is currently pursuing a Doctorate in Information Technology at La Consolacion University Philippines. His academic and research interests focus on Software Engineering and Machine Learning.

Dr. Jonilo C. Mababa is the current President of the Philippine Society of Information Technology Educators (PSITE) – Central Luzon Chapter (2022–2025). He is a dedicated graduate school lecturer at Holy Angel University, La Consolacion University Philippines, and Systems Plus College Foundation. With a strong background in academic leadership, he previously served as the Dean of AMA Computer College – Angeles. His work focuses on advancing IT education and leadership in higher education institutions.