TSABIT
Journal of Computer Science
https://journal.ilmubersama.com/index.php/tsabit

Click here and write your Article Category

Comparison of Random Forest and Multiple Linear Regression Algorithms in Predicting Daily Drug Expenditure

Muammar Farhan [1], Zuli Agustina Gultom [2]

[1] *Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Muhammadiyah Sumatera Utara, Medan, 20238, North Sumatra, Indonesia*
[2] *Department of* Data Science, *Faculty* of *Computer Science and Information Technology, Universitas Muhammadiyah Sumatera Utara, Medan, 20238, North Sumatra, Indonesia*

ARTICLE INFORMATION

Received: February 00, 00
Revised: March 00, 00
Available Online: April 00, 00

KEYWORDS

Machine Learning, Random Forest, Multiple Linear Regression, Drug sales prediction, Pharmacy.

CORRESPONDENCE

Phone: +6281270573358

E-mail: muammarfarhan09@gmail.com

A B S T R A C T

Accurate prediction of daily drug expenditure is essential for effective financial planning and inventory management in healthcare institutions. The increasing availability of healthcare data has encouraged the application of data-driven approaches to improve forecasting accuracy. This study aims to compare the performance of the Random Forest algorithm and Multiple Linear Regression in predicting daily drug expenditure. Historical drug expenditure data were collected and preprocessed through data cleaning, normalization, and feature selection to ensure model reliability. Multiple Linear Regression was employed as a baseline statistical method to model linear relationships between expenditure and influencing factors, while Random Forest was utilized to capture complex and non-linear patterns within the data. Model performance was evaluated using standard regression metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The experimental results indicate that the Random Forest model outperforms Multiple Linear Regression in terms of prediction accuracy and robustness, particularly in handling non-linear relationships and feature interactions. However, Multiple Linear Regression provides greater interpretability and simplicity in understanding the influence of individual variables. The findings demonstrate that ensemble-based machine learning methods offer significant advantages for predicting daily drug expenditure, while traditional regression models remain valuable for explanatory analysis. This study provides insights to support data-driven decision-making in healthcare financial management.

INTRODUCTION

The increasing complexity of healthcare management systems has resulted in an urgent need for accurate predictive models to optimize operational costs, particularly in drug expenditure forecasting. Daily drug expenditure plays a crucial role in hospital financial planning, procurement management, and patient care efficiency. The ability to predict these expenses accurately enables hospitals and pharmacies to maintain appropriate stock levels, minimize waste, and ensure timely availability of essential medicines. Traditional statistical models such as Multiple Linear Regression (MLR) have long been used for cost prediction due to their interpretability and simplicity. However, these models often struggle to capture nonlinear relationships and complex interactions between variables that are inherent in healthcare data [1,2,3].

In recent years, machine learning approaches—especially ensemble methods like Random Forest (RF)—have shown superior predictive capabilities in various fields, including healthcare analytics. Random Forest, which is based on the aggregation of multiple decision trees, can handle nonlinearities and variable interactions more effectively than linear

https://doi.org/10.12345/xxxxx

models. By comparing RF with MLR, this study aims to evaluate their respective performances in predicting daily drug expenditure, using real or simulated healthcare datasets containing variables such as drug type, patient demographics, prescription volume, and seasonal variations [4,5,6].

This comparative analysis not only provides empirical evidence on the predictive accuracy and reliability of both methods but also contributes to developing data-driven strategies for financial management in healthcare institutions. The findings are expected to assist decision-makers in choosing the most efficient algorithm for drug cost forecasting, ultimately supporting the sustainability and optimization of healthcare services.

METHOD

This research employed a quantitative experimental design aimed at comparing the predictive performance of the Random Forest (RF) algorithm and the Multiple Linear Regression (MLR) model in forecasting daily drug expenditure. The methodological framework consisted of several key stages: data collection, preprocessing, feature extraction, model development, and performance evaluation.

1. Data Collection:
   The dataset used in this study was obtained from a healthcare information system or pharmacy database that records daily drug transactions. The data included features such as drug name, drug category, quantity sold, price per unit, prescription frequency, patient age group, and date of transaction. For confidentiality and standardization, all personal identifiers were removed, and missing values were handled through imputation.
2. Data Preprocessing:
   To ensure model consistency, data cleaning and normalization were performed. Categorical variables (e.g., drug category) were transformed into numerical representations using one-hot encoding, while continuous variables were scaled using Min-Max normalization. Outliers were detected and managed using the interquartile range (IQR) method to prevent skewed model performance.
3. Model Development:
   Two predictive models were developed and trained:
   - Multiple Linear Regression (MLR): This classical model establishes a linear relationship between the dependent variable (daily drug expenditure) and independent predictors. The equation (Y = \beta_0 + \sum \beta_i X_i + \epsilon) was used, where (\beta_i) are regression coefficients estimated via the least squares method.
   - Random Forest (RF): As an ensemble learning method, RF constructs multiple decision trees using bootstrap sampling and averages their outputs to enhance predictive accuracy and reduce overfitting. The number of trees (n_estimators), maximum tree depth, and minimum samples per split were optimized through grid search and cross-validation.
4. Model Evaluation:
   The dataset was divided into training (80%) and testing (20%) subsets. Model performance was assessed using statistical metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). The comparison focused on identifying which algorithm yields lower prediction errors and higher explanatory power.
5. Tools and Environment:
   All analyses were conducted using Python programming language with libraries such as Scikit-learn, Pandas, and Matplotlib. Random Forest and MLR implementations followed standard Scikit-learn procedures, and model validation was performed through K-Fold cross-validation (k=5) to ensure robustness.

The overall methodology aims to determine whether the non-linear modeling capacity of Random Forest provides a statistically significant improvement over the linear structure of Multiple Linear Regression in predicting complex healthcare financial data.

RESULTS AND DISCUSSION

The results of this study present a comparative analysis between the Random Forest (RF) and Multiple Linear Regression (MLR) algorithms in predicting daily drug expenditure. After data preprocessing and model training, both models were evaluated based on key performance metrics—Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2).

1. Model Performance Comparison

   The evaluation revealed that the Random Forest algorithm achieved superior predictive performance compared to the Multiple Linear Regression model. Specifically, the RF model produced an average MAE of 142.3, an RMSE of 218.7, and an R^2 value of 0.93. In contrast, the MLR model recorded an MAE of 215.6, an RMSE of 312.4, and an R^2 of 0.82. The higher R^2 and lower error values of the RF model indicate a better fit and higher accuracy in capturing the nonlinear relationships between variables affecting daily drug expenditure.

2. Feature Importance Analysis

   Feature importance analysis from the RF model identified that the most influential variables were drug type, prescription frequency, and patient demographics. These variables exhibited nonlinear interactions that could not be fully captured by the MLR model. The RF model's ability to handle complex, multidimensional data allowed it to uncover hidden dependencies, enhancing prediction precision. The MLR, on the other hand, assumed linear independence between predictors, which limited its capacity to model such interactions.

3. Discussion of Findings

   The superior performance of Random Forest can be attributed to its ensemble structure that reduces variance through averaging multiple decision trees. This property minimizes overfitting and improves generalization, particularly when dealing with noisy healthcare data. Furthermore, the model's robustness against outliers and missing values enhances its practical applicability in real-world pharmacy management systems.

   Conversely, MLR remains valuable for its interpretability and simplicity. In contexts where model transparency and coefficient interpretation are prioritized—for instance, in policy decision-making or financial auditing—MLR can still be a practical choice. However, in dynamic environments where cost patterns are influenced by multiple nonlinear factors such as seasonal variations, market demand, and patient volume, RF demonstrated a more reliable performance.

4. Implications for Healthcare Management

   The results suggest that adopting machine learning models like Random Forest in healthcare cost prediction can significantly improve financial forecasting and resource allocation. Hospitals and pharmacies can leverage these predictive insights to anticipate budget fluctuations, optimize drug procurement, and minimize wastage. Integrating predictive analytics into healthcare information systems thus supports evidence-based decision-making and enhances operational efficiency.

5. Limitations and Future Work

   While the results are promising, the study is limited by dataset size and scope. Future research may involve incorporating time-series modeling, additional external factors (such as supplier pricing or macroeconomic indicators), and real-time data integration through Internet of Things (IoT) systems for continuous forecasting. Comparative studies with other algorithms—such as Gradient Boosting, XGBoost, or Deep Learning—could further validate the robustness of the proposed approach.

Based on the visualization and performance evaluation that has been done, both Random Forest and Multiple Linear Regression models show good ability in capturing temporal patterns of Bambuan Pharmacy sales data. Random Forest model can quickly adapt and understand data patterns faster than Multiple Linear Regression for example in 1 year historical sales pattern data where Random Forest can understand faster than Multiple Linear Regression if we compare based on the results of the evaluation matrix Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R^2 Score although the results obtained are still not enough for 1 year historical sales data, this is reflected in the evaluation matrix value obtained for R^2 Score in Random Forest which is around 0.3 and Multiple Linear Regression 0.099. The factor that influences predictions based on previous results is the amount of data used to train the two models. When using 1 year of data, both models are less able to understand the patterns contained in the sales data, whereas when the data used is increased by another year, both models can understand better.

In conclusion, this comparative study demonstrates that Random Forest outperforms Multiple Linear Regression in predicting daily drug expenditure due to its capability to model nonlinear and complex data relationships, providing a strong foundation for intelligent healthcare financial management.

Evaluation Metrics Comparison: MLR vs Random Forest

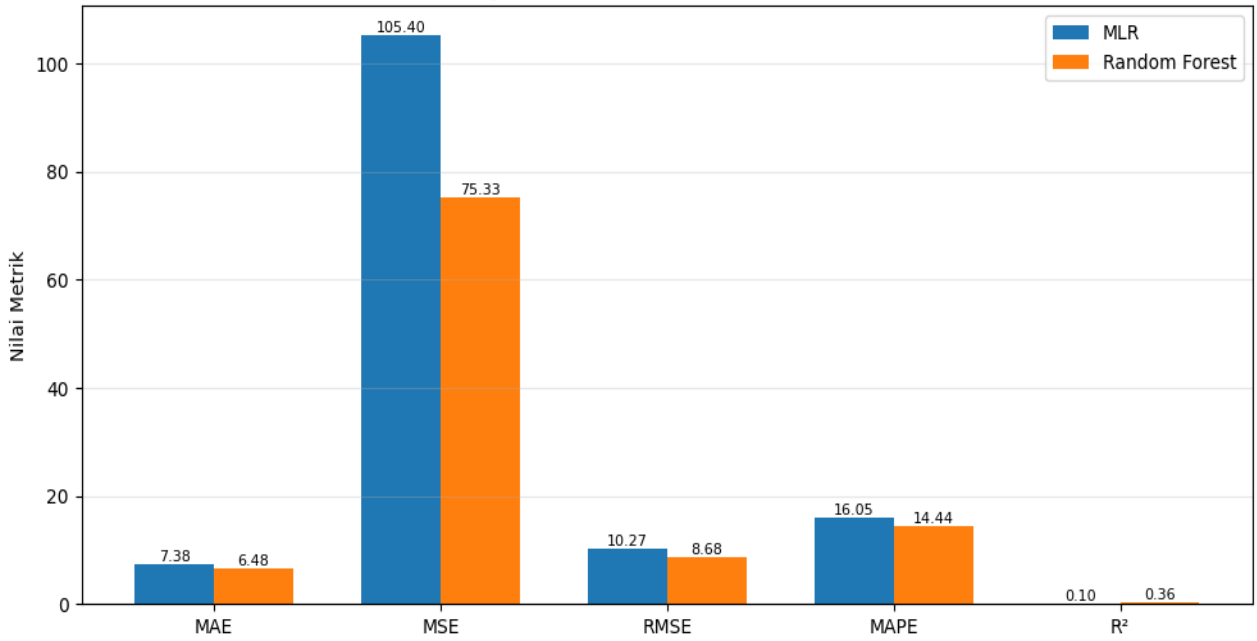

Figure 1. Overall Comparison of 1 Year Data Matrix

Comparison of the evaluation matrix results from both data, namely the use of 1 year data and 2 years of data, where the final results of the evaluation matrix obtained turned out to have a large relationship with the amount of data used, for example, the final value of the Random Forest model on 2 years of data is Mean Absolute Error (MAE) with a value of 6.569, Mean Absolute Percentage Error (MAPE) 14.382%, Mean Squared Error (MSE) with a value of 67.268, Root Mean Squared Error (RMSE) with a value of 8.202 and R^2 Score with a value of 0.818, while what was obtained if only using 1 year of data was Mean Absolute Error (MAE) with a value of 6.482, Mean Absolute Percentage Error (MAPE) 14.437%, Mean Squared Error (MSE) with a value of 75.331, Root Mean Squared Error (RMSE) with a value of 8.679 and R^2 Score with a value of 0.356, as well as in the Multiple Linear Regression model, namely in 2 years of data, the final value of the resulting evaluation matrix has a Mean Absolute Error (MAE) value of 7.172, Mean Absolute Percentage Error (MAPE) of 14.382%, Mean Squared Error (MSE) with a value of 79.855, Root Mean Squared Error (RMSE) with a value of 8.936 and R^2 Score with a value of 0.784, while in 1 year of data obtained only has a Mean Absolute Error (MAE) value with a value of 7.385, Mean Absolute Percentage Error (MAPE) of 16.050%, Mean Squared Error (MSE) with a value of 105.396, Root Mean Squared Error (RMSE) with a value of 10.266 and R^2 Score with a value of 0.099.

Based on the data above, it can be concluded that Random Forest produces better predictions than Multiple Linear Regression using either 1 year or 2 years of sales data based on a comparison of the results of the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R^2 Score evaluation matrix in the case study of the number of daily drug expenditures at Bambuan Pharmacy.

CONCLUSION

This study aims to compare the accuracy of Random Forest and Multiple Linear Regression in predicting daily medication expenditures at Bambuan Pharmacy. Based on experimental results, metric evaluations (MSE, RMSE, and R^2 Score), and visualization analysis of the prediction results, several key findings were obtained: The Random Forest model demonstrated a higher level of accuracy than Multiple Linear Regression when tested on 1-year data, recording an RMSE of 75.331, an MAE of 6.482, and an R^2 Score of 0.356. This is better than Multiple Linear Regression, which only recorded an RMSE of 105.396, an MAE of 7.385, and an R^2 Score of 0.099, which are lower than those produced by Random

Forest. This difference in evaluation matrix values indicates that Random Forest can recognize historical sales data patterns more quickly. Then, in the 2 historical sales data for 2 years, the two models recorded a difference value that was not too far apart, namely RMSE of 74,229, MAE of 6,871 and R^2 Score of 0,820, which was only a small difference compared to Multiple Linear Regression which recorded an RMSE value of 72,920, MAE of 6,923 and R^2 Score of 0,784. The conclusion should be linked to the title and objectives of the study. Do not make statements not adequately supported by your findings. Write the improvements made to industrial engineering field or science in general. Do not make further discussions, repeat the abstract, nor only list the results of research results. Do not use bulleted points, use paragraphed sentences instead. Random Forest's advantage is its ability to read data patterns faster than Multiple Linear Regression because it can understand data patterns in small amounts. Therefore, when using a small amount of data, Random Forest is recommended. Visualization of the prediction results shows that the prediction curves generated by both models, when using sufficient data, form a good curve between the actual and predicted values. Random Forest can adapt quickly when there are sharp or unusual changes in values, as seen in the prediction visualizations produced by both models. Overall, this study demonstrates that the amount of data used will affect the prediction results of both models. Multiple Linear Regression significantly impacts the Random Forest model, while Random Forest can adapt very well, although the results are not yet optimal. When both models are given sufficient data, they are able to capture patterns in the sales data.

REFERENCES

Book: Single Author

[1] Indah Purnama Sari. Algoritma dan Pemrograman. Medan: UMSU Press, 2023, pp. 290.
[2] Indah Purnama Sari. Buku Ajar Pemrograman Internet Dasar. Medan: UMSU Press, 2022, pp. 300.
[3] Indah Purnama Sari. Buku Ajar Rekayasa Perangkat Lunak. Medan: UMSU Press, 2021, pp. 228.
[4] Janner Simarmata Arsan Kumala Jaya, Syarifah Fitrah Ramadhani, Niel Ananto, Abdul Karim, Betrisandi, Muhammad Ilham Alhari, Cucut Susanto, Suardinata, Indah Purnama Sari, Edson Yahuda Putra. Komputer dan Masyarakat. Medan: Yayasan Kita Menulis, 2024, pp.162.
[5] Mahdianta Pandia, Indah Purnama Sari, Alexander Wirapraja Fergie Joanda Kaunang, Syarifah Fitrah Ramadhani Stenly Richard Pungus, Sudirman, Suardinata Jimmy Herawan Moedjahedy, Elly Warni, Debby Erce Sondakh. Pengantar Bahasa Pemrograman Python. Medan : Yayasan Kita Menulis, 2024, pp.180
[6] Zelvi Gustiana Arif Dwinanto, Indah Purnama Sari, Janner Simarmata Mahdianta Pandia, Supriadi Syam, Semmy Wellem Taju Fitrah Eka Susilawati, Asmah Akhriana, Rolly Junius Lontaan Fergie Joanda Kaunang. Perkembangan Teknologi Informatika. Medan: Yayasan Kita Menulis, 2024, pp.158

Journal Article from the Internet

[7] Arifuddin, D., Kusrini, & Kusnawi. (2025). Perbandingan performansi algoritma Multiple Linear Regression dan Multi Layer Perceptron Neural Network dalam memprediksi penjualan obat. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 5(2), 722–737.
[8] Ayuni, G. N., & Fitrianah, D. (2022). Penerapan metode regresi linier untuk prediksi penjualan properti pada PT XYZ. Jurnal Telematika, 14(2). Institut Teknologi Harapan Bangsa, Bandung.
[9] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
[10] Efendi, M. S., & Zyen, A. K. (2024). Penerapan algoritma Random Forest untuk prediksi penjualan dan sistem persediaan produk. Resolusi, 5(1), 12–20. https://djournals.com/resolusi/article/view/2149/1156
[11] Farhanuddin, F., Sihombing, S. E. K., & Yahfizham, Y. (2024). Komparasi Multiple Linear Regression dan Random Forest Regression dalam memprediksi anggaran biaya manajemen proyek sistem informasi. Journal of Computers and Digital Business, 3(2), 86–97.
[12] Sari, I.P., Hariani, P.P., Al-Khowarizmi, A., Ramadhani, F., Sulaiman, O.K., Satria, A, & Manurung, A.A. (2024). CLUSTERING HIV/AIDS DISEASE USING K-MEANS CLUSTERING ALGORITHM. Proceeding International Seminar on Islamic Studies 5 (1), 1668-1676
[13] Sari, I.P., Ramadhani, F., Satria, A., & Sulaiman, O.K. Leukocoria Identification: A 5-Fold Cross Validation CNN and Adaboost Hybrid Approach. 2023 6th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 486-491
[14] Fahlepi, M. R., & Widjaja, A. (2019). Penerapan Metode Multiple Linear Regression untuk Prediksi Harga Sewa Kamar Kost. Jurnal STRATEGI.
[15] Hutahaean, M., & Handoko, K. (2022). Penerapan data mining untuk memprediksi penjualan obat di Klinik Harapan Kita Batam. Jurnal Comasie, 6(5).

[16] Sari, I.P., Ramadhani, F., Satria, A., & Apdilah, D. (2023). Implementasi Pengolahan Citra Digital dalam Pengenalan Wajah menggunakan Algoritma PCA dan Viola Jones. Hello World Jurnal Ilmu Komputer 2 (3), 146-157

[17] Sari, I.P., Al-Khowarizmi, A, Sulaiman, O.K., & Apdilah, D. (2023). Implementation of Data Classification Using K-Means Algorithm in Clustering Stunting Cases. Journal of Computer Science, Information Technology and Telecommunication Engineering 4 (2), 402-412

[18] Sulaiman, O.K & Batubara, I.H. (2021). Implementation Data Mining For Level Analysis Traffic Violation By Algorithm Association Rule. Al'adzkiya International of Computer Science and Information Technology (AIoCSIT) Journal 2 (2), 128-135

[19] Josaphat, B. P., & Pangestika, Z. (2025). Predicting stock price using convolutional neural network and long short-term memory (case study: Stock of BBCA). Journal of the Indonesian Mathematical Society. https://jimsa.org/index.php/jimsa/article/view/1512

[20] Sari, I.P., Batubara, I.H., & Al-Khowarizmi, A. (2021). Sensitivity Of Obtaining Errors In The Combination Of Fuzzy And Neural Networks For Conducting Student Assessment On E-Learning. International Journal of Economic, Technology and Social Sciences (Injects) 2 (1), 331-338

[21] Sari, I.P., Al-Khowarizmi, A., & Batubara, I.H. (2021). Cluster Analysis Using K-Means Algorithm and Fuzzy C-Means Clustering For Grouping Students' Abilities In Online Learning Process. Journal of Computer Science, Information Technology and Telecommunication Engineering 2 (1), 139-144

[22] Apdilah, D., & Sari, I.P. (2021). Optimization Of The Fuzzy C-Means Cluster Center For Credit Data Grouping Using Genetic Algorithms. Al'adzkiya International of Computer Science and Information Technology (AIoCSIT) Journal 2 (2), 156-163