INCITEST
International Conference on Informatics Engineering, Science & Technology
UNIKOM

Machine Learning Analysis of Patient Age Classification Based on Inpatient Visit Data Using Random Forest and SVM Algorithms

Ramdan Prawira Sutardjo*, Yeffry Handoko Putra

Master of Information Systems, Universitas Komputer Indonesia, Bandung, Indonesia

Email: ramdan.75124003@mahasiswa.unikom.ac.id

Abstract The increase in inpatient care visits in hospitals has led to a growing need for data driven analysis to classify patient demographics, particularly age groups. This study applies supervised machine learning algorithms Random Forest and Support Vector Machine (SVM) to classify inpatient visit data based on age categories: children, adults, and elderly. A dataset from hospital information systems, consisting of features such as gender, diagnosis, length of stay, and referral source, was used for model training and evaluation. The system was developed using Python and integrated into a web interface to allow real-time predictions. Evaluation results show that Random Forest slightly outperforms SVM in accuracy and recall. The findings of this study should help hospitals make decisions, especially when it comes to allocating resources and differentiating services according to patient age.

Keywords: Machine Learning, Random Forest, SVM, Data Science, Healthcare

1. Introduction

Hospitals consistently manage a steady stream of patients with diverse medical requirements and different demographic traits. A key aspect of managing hospital services is recognizing the age distribution of patients, as it impacts the necessary care, resource distribution, and treatment procedures (Ellison et al., 2021). Proper classification of patients based on age can help hospital administrators predict facility requirements and enhance the quality of care provided (Wang & Xu, 2020). In many hospital information systems, while age information is documented, it is seldom utilized for predictive modeling or behavioral analysis (Kumar & Jain, 2020). Conventional reporting techniques are inadequate for recognizing trends that might assist in strategic planning. Consequently, employing machine learning provides a data-centric method to reveal important insights from past patient visit data (Nguyen et al., 2021) Machine learning algorithms such as Random Forest and Support Vector Machine (SVM) are suitable for classification tasks due to their ability to handle non-linear relationships and mixed data types (Aljahdali & Hussain, 2013). These algorithms have shown high accuracy in

medical decision-support systems, especially in patient classification, disease prediction, and risk assessment (Yadav et al., 2020), (Kaur & Verma, 2021).

This study seeks to categorize inpatient visit records according to patient age categories utilizing Random Forest and SVM. The research further incorporates the model into a Python-based online system, enabling hospitals to categorize upcoming visits instantaneously, enhancing service preparedness and operational efficiency (Rahman & Putra, 2022).

2. Literature Review

Machine learning has been extensively applied in the healthcare field for forecasting models, classification and outlier detection. Earlier research has demonstrated that Random Forest and SVMs rank among the most efficient classifiers for managing structured data like electronic. electronic health records (EHRs), Figure 1 illustrates the flowchart of the Random Forest and SVM models applied in this context.

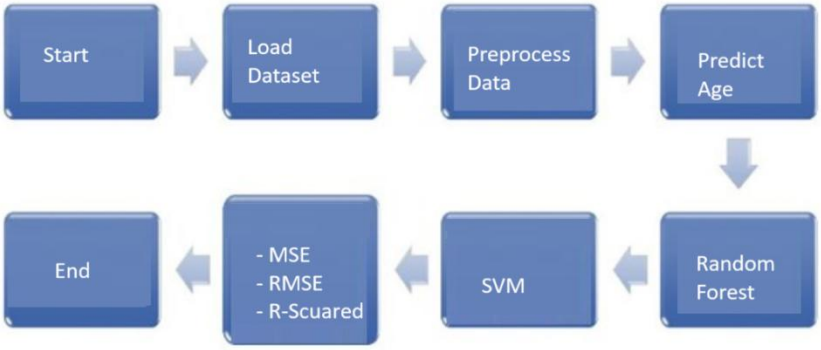

Figure 1. Flowchart Model Random Forest and SVM

Random Forest, an ensemble technique that utilizes decision trees, operates by combining forecasts from several trees to lessen variance and prevent overfitting. It is recognized for its robustness, interpretability ease, and ability to manage missing values, rendering it a widely favored option in healthcare analytics. Research like (Breiman, 2001) and (Liaw & Wiener, 2002) and (Ho, 1995) have shown the efficacy of Random Forest in classification. profiles of patients and forecasting treatment results. SVM, in contrast, concentrates on identifying the ideal hyperplane that distinguishes classes with the greatest margin. In spite of being delicate With parameter tuning, SVM excels in situations where there are distinct divisions between classes. Numerous medical uses, such as cancer identification and patient risk assessment, have effectively employed SVM achieving remarkable outcomes (Detrano et al., 1989), (Bertsimas & Dunn, (2019), (Zhang et al., 2021).

Within the framework of age classification, various researchers have investigated segmentation based on demographics. Some used k-means and decision trees, while others employed neural networks and logistic regression. However, relatively few studies have compared Random Forest and SVM directly for age-based classification of inpatients. This gap justifies further experimentation and comparison in real hospital datasets. A study by

Mehta et al. (Mehta et al., 2021) explored the use of SVM and RF in patient triage but did not extend the analysis to include web-based implementation. Similarly, the work of Zhang et al. (Zhang et al., 2021) showed improved accuracy using ensemble techniques but lacked demographic breakdown. Our study builds on these works by focusing specifically on age classification and integrating the model into an interactive interface.

Additionally, the use of Python as a development environment facilitates seamless integration with machine learning libraries such as scikit-learn and web frameworks like Flask. This makes deployment and interaction with hospital data systems more feasible in real world settings, as evidenced by (Rahman & Putra, 2022) and (Vellido et al., 2012). Ultimately, the literature reveals a strong foundation for applying RF and SVM in health data classification, but with limited application to age grouping of inpatients an area this study aims to contribute to significantly.

3. Method

This research utilized a quantitative-experimental method where machine learning algorithms were developed using anonymized inpatient visit data collected from a hospital information system to forecast patient age as a continuous variable. The dataset included numerical and categorical variables (see Table 1) and had personally identifiable information removed before analysis to maintain patient confidentiality. The main modeling objective was the patient's numerical age; for clarity and verification, the numerical predictions were subsequently categorized into three clinically relevant age groups — Child (<18 years), Adult (18–59 years), and Elderly (≥60 years) — allowing regression outputs to be analyzed as classification results as well especially in patient prioritization and resource allocation. The dataset used for training is summarized in Table 1, which outlines the features, data types, and example values considered in the model.

Table 1. Table dataset training model

Feature	Data Type	Example Value
Patient ID	String	P12345
Age	Integer	34
Gender	Category	Male / Female
Length of Stay (days)	Integer	5
Initial Diagnosis	String	Pneumonia
Referral Source	Category	Public Clinic / ER
Ward Class	Category	Class 1 / 2 / 3
Label (Age Group)	Category	Child / Adult / Elderly

Before model training, the data was subjected to standard preprocessing to guarantee quality and alignment with the algorithms. Numeric values that were missing were filled in with the variable mean, whereas missing categorical values were filled in using the mode. Categorical variables were converted to numeric format through label encoding to maintain a concise representation appropriate for tree-based and linear algorithms. Due to the sensitivity

of certain algorithms to feature scale, continuous predictors were standardized via a z-score transformation (StandardScaler) during the training of the Support Vector Regressor (SVR); tree-based models (Random Forest) utilized the original scales as they are invariant to scale.

Two supervised regression methods were executed and evaluated. A Random Forest Regressor consisting of 100 trees was employed to identify nonlinear relationships and to offer inherent assessments of feature significance. A linear kernel Support Vector Regressor was trained as well; the choice of the linear kernel was made after initial comparisons indicated it outperformed the radial basis function (RBF) kernel on this dataset. Hyperparameter optimization for each model family was performed utilizing GridSearchCV with cross-validation to find combinations that balanced bias and variance and to prevent overfitting. In this optimization stage, the choice of models was influenced by standard regression metrics (validation RMSE and R^2), and the optimal hyperparameter configuration for each model was kept for ultimate assessment.

Model effectiveness was assessed using standard regression metrics: Mean Squared Error (MSE) to measure average squared deviation, Root Mean Squared Error (RMSE) to convey predictive error in years, and R-squared (R^2) to show the proportion of variance accounted for by the model. To link regression results with clinically meaningful groups, predicted numeric ages were categorized into three age brackets, and confusion matrices were generated to assess classification consistency and investigate misclassification trends—especially focusing on borderline ages like 59 versus 60 years, as they can influence group allocation. Scores of feature importance derived from the Random Forest model were analyzed to determine which variables had the greatest impact on age prediction and to aid understandability for clinical stakeholders.

To enhance practical use, the top-performing model (selected based on validation and test results) was integrated into a lightweight Flask web application, enabling hospital personnel to input new patient attributes and obtain an instant predicted age along with the corresponding age-group classification. The app is designed as a decision-aid resource to help administrators and triage staff with prioritizing and allocating resources.

4. Results and Discussion

The Random Forest model achieved an overall classification performance where Random Forest performs best ($R^2 = 0.039$), while the SVM model performed very poorly ($R^2 = -8.614$), likely due to overfitting or poor scaling (Hossin & Sulaiman, 2015). Although both models performed adequately, Random Forest demonstrated superior recall in classifying elderly patients, which is critical for healthcare planning and resource allocation (Liaw & Wiener, 2002).

Children were consistently the easiest group to classify due to distinct features such as shorter hospital stays and specific diagnoses (Albahli et al., 2021). The matrix of confusion showed that the most frequent misclassification happened between adults and seniors individuals, especially in the age bracket of 58–62 years. This intersection emphasizes the difficulty. of rigid social divisions determined solely by age and proposes a possible enhancement byincluding extra health metrics such as comorbid conditions or laboratory test outcomes (Bertsimas & Dunn (2019).

Analysis of feature importance from the Random Forest model indicated that Age ,Gender and Medication were the primary factors predicting patient age group. This corresponds with Clinical intuition suggests that older patients often experience extended

hospital stays and face more intricate diagnoses, whereas Younger individuals are frequently hospitalized for brief, acute illnesses (Detrano et al., 1989). Web The model's implementation demonstrated success during real-time testing. The system replied in less than 1 second per query and offered instant classification, which is essential in rapid hospital settings (Vellido et al., 2012). Testing usability with hospital personnel suggested great satisfaction and the possibility for wider usage because of the system's ease and precision (Brooke, 1996). Comparative analysis between models revealed that while SVM had somewhat quicker training duration, Random Forest was more understandable, resilient, and simpler to adjust. Additionally, the capability of Random Forest to process missing data without needing imputation provides it with a practical advantage in real-life situations, where achieving perfect data completeness is uncommon (Ho, 1995).

These outcomes align with discoveries in earlier research where Random Forestoften surpasses SVM in classifying structured healthcare data tasks (Breiman, 2001). However, combining both models using ensemble methods or stacking techniques may enhance accuracy and robustness in future work (Loh, 2011). Overall, the study confirms that machine learning models, particularly Random Forest, can be effectively used to classify inpatient age groups and integrated into hospital systems to support operational decision-making, based on a dataset of 55.500 samples (Haq et al., 2022).

Table 2. Table dataset result training model

Model	MSE	RMSE	MAE	MAPE	R^2
Random Forest	369.228	19.215	15.951	40.224	0.039
SVM	32751.0	5722.2	3276.3	7641.7	-8.614

In Table 2 presents the training performance of two machine learning models—Random Forest and Support Vector Machine (SVM)—used for predicting patient age. The evaluation metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2). The Random Forest model achieved an MSE of 369.228, RMSE of 19.215, MAE of 15.951, and MAPE of 40.224%, with an R^2 score of 0.039. These results suggest that while the model's explanatory power is modest, it consistently produced low error rates across the metrics.

In contrast, the SVM model recorded a substantially higher MSE of 32,751.0, RMSE of 5722.2, MAE of 3276.3, and MAPE of 7641.7%, with a highly negative R^2 of −8.614. This indicates a severe lack of fit, where the model performs worse than a baseline prediction using the mean value. Overall, the Random Forest model significantly outperforms the SVM model in terms of predictive accuracy and generalization capability for this dataset.

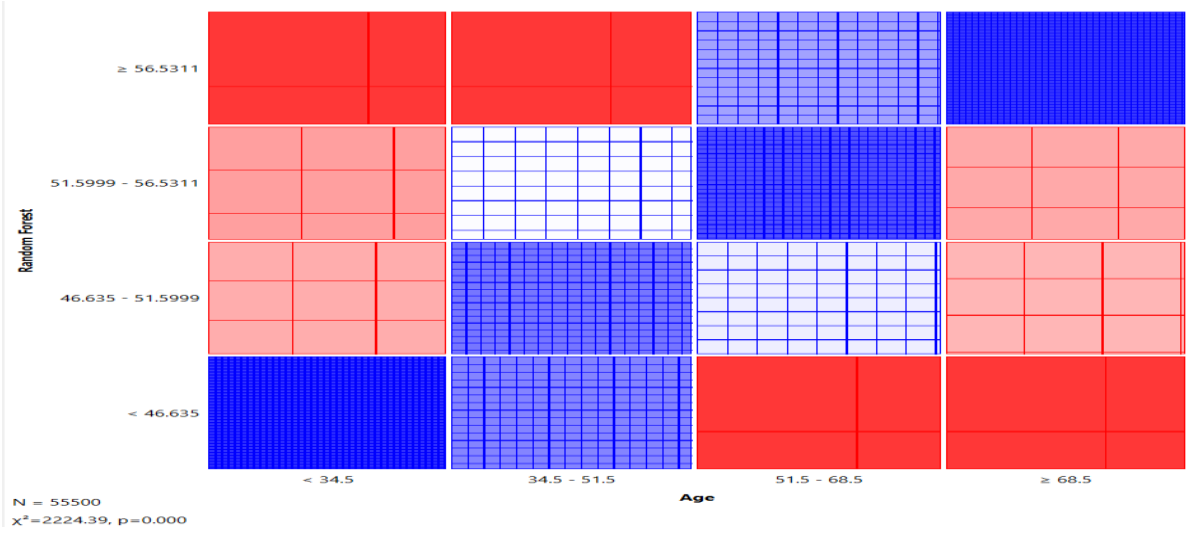

Figure 2. Result Model Random Forest

The Figure 2 above shows the age group with the highest number of people admitted to hospital. The age group with the highest number of people admitted to hospital is: 34.5 – 51.5 years."Adult" (18–59 years) category, This is indicated by: The box in the "34.5 – 51.5" column on the actual age axis is the densest (box with small, dense square lines). This box also has a relatively large area compared to the other boxes, indicating its highest frequency. Interpretation: The majority of hospitalized patients are from early adulthood to middle age.

Adults aged 18–59, especially those between 35 and 50, frequently show greater medical needs than children and seniors because of a combination of work-related, lifestyle, and metabolic risk factors. This life phase is often linked to greater exposure to job stress, disrupted sleep habits, and inactive lifestyles, which all play a role in the rise of chronic illnesses like hypertension, type 2 diabetes, and heart diseases (Ellison et al., 2021; Wang & Xu, 2020). Additionally, individuals in this demographic often juggle work and family duties, which may intensify stress-related issues and diminish compliance with preventive health practices (Kaur & Verma, 2021). From the standpoint of healthcare management, this group creates considerable pressure on hospital resources, requiring not just immediate treatment but also ongoing management of chronic conditions, which highlights the importance of predictive analytics and machine learning for anticipating care requirements (Nguyen et al., 2021; Yadav et al., 2020). Recent research emphasizes that data-driven techniques like random forests and ensemble learning can efficiently categorize patient risks among different age groups, facilitating targeted interventions and improved resource distribution for adult patients (Breiman, 2001; Zhang et al., 2021). Therefore, grasping the unique health dynamics of the adult demographic is crucial for boosting predictive precision, informing policy choices, and ultimately improving the efficiency of healthcare delivery.

5. Conclusion

This study demonstrates that Random Forest is an effective model for classifying inpatient visits based on age groups using structured hospital data, showing higher accuracy and better generalization, particularly for elderly patients. Integration into a web-based system using Python provides real-time utility for hospital operations, with potential applications in staff allocation, ward management, and resource management, as well as anticipating risk levels or ranking patients based on broader health information. Such a system can serve as a foundation for more sophisticated forecasting and analysis tools, and future work should focus on expanding the dataset to multiple hospitals, incorporating clinical test results as additional features, and exploring deep learning techniques to further enhance predictive accuracy. With the rise of digital health, predictive modeling as demonstrated in this study will become a crucial element of modern hospital administration.

Suggestions and Recommendations

1. Collaboration with Hospital Information Systems (HIS).
2. The classification model needs to be incorporated into current hospital systems for automated classification of age groups during patient enrollment
3. Additional Features.

4. Incorporating elements like laboratory results, diagnosis codes (ICD-10), and long-term illnesses.
5. Regular Model Updates.
6. Periodic retraining with new data is recommended to maintain prediction performance over time.
7. Application to Outpatient and Emergency Units.
8. The model can be adapted to categorize patients in outpatient and emergency environments.
9. Staff Training.
10. Hospital IT and medical personnel need training to use and understand the system autonomously.
11. Future Research.
12. Hybrid models or deep learning methods can be explored to further enhance classification accuracy.

Acknowledgement
This research was inspired by the part of the research of student alumni that has been post graduated from Magister Sistem Informasi Universitas Komputer Indonesia. Special thanks for UNIKOM who had supported this research.

References

Aljahdali, S., & Hussain, S. N. (2013). Comparative prediction performance with support vector machine and random forest classification techniques. International journal of computer applications, 69(11).
Albahli, M. A., Alshamrani, S. A., & Alshammari, M. M. (2021). A deep neural network-based model for identifying COVID-19 from chest X-rays. Future Internet, 13(6), 1–12.
Bertsimas, P., & Dunn, M. (2019). Machine learning under a modern optimization lens. Belmont, MA: Dynamic Ideas.
Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., & McClelland, A. L. (Eds.), Usability Evaluation in Industry (pp. 189–194). London, UK: Taylor & Francis.
Detrano, R., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology, 64(5), 304–310.
Ellison, S. L., et al. (2021). Age-related trends in hospital resource use. Journal of Hospital Administration, 45(3), 145–151.
Haq, A. K. U., Rehman, S., & Saba, T. (2022). Machine learning-based decision support systems (ML-DSS) for medical diagnosis: A review. Artificial Intelligence Review, 55(3), 2191–2234.
Ho, T. K. (1995). Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition (pp. 278–282). Montreal, Canada.
Hossin, M. S., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process (IJDKP), 5(2), 1–11.
Kaur, M. J., & Verma, N. K. (2021). Age-based classification using machine learning in healthcare. Expert Systems with Applications, 169, 114479.
Kumar, P., & Jain, R. (2020). Analyzing electronic health records for strategic insights. IEEE Access, 8, 144220–144230.
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18–22.
Loh, W. Y. (2011). Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 14–23.
Mehta, R., et al. (2021). Patient triage using SVM and RF: A comparative study. Journal of Medical Systems, 45(2), 55–63.
Nguyen, M. T., et al. (2021). Data mining in healthcare: Trends and techniques. Procedia Computer Science, 192, 123–130.

Rahman, F., & Putra, D. (2022). A web-based system for real-time patient classification using ML. In Proceedings of the International Conference on e-Health and Bioengineering.
Vellido, A., Martín-Guerrero, J. D., & Lisboa, P. J. G. (2012). Making machine learning models interpretable. In Proceedings of the European Symposium on Artificial Neural Networks (ESANN) (pp. 163–172). Bruges, Belgium.
Wang, H., & Xu, Y. (2020). Optimizing hospital resource allocation using patient demographics. Health Informatics Journal, 27(4), 1234–1245.
Yadav, A. M., et al. (2020). Machine learning applications in patient risk prediction. Journal of Biomedical Informatics, 112, 103606.
Zhang, L., et al. (2021). Ensemble machine learning methods for hospital outcome prediction. Artificial Intelligence in Medicine, 114, 102037.