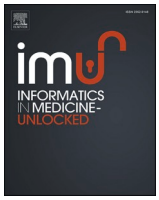
Informatics in Medicine Unlocked 47 (2024) 101499

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Comparison of nine machine learning regression models in predicting hospital length of stay for patients admitted to a general medicine department

Addisu Jember Zeleke [a,*], Pierpaolo Palumbo [a], Paolo Tubertini [b], Rossella Miglio [c], Lorenzo Chiari [a,d]

[a] Department of Electrical, Electronic, and Information Engineering Guglielmo Marconi, University of Bologna, 40126, Bologna, Italy
[b] Enterprise Information Systems for Integrated Care and Research Data Management, IRCCS Azienda Ospedaliero- Universitaria di Bologna, 40138, Bologna, Italy
[c] Department of Statistical Sciences, University of Bologna, 40126, Bologna, Italy
[d] Health Sciences and Technologies Interdepartmental Center for Industrial Research (CIRI SDV), University of Bologna, 40126, Bologna, Italy

ARTICLE INFO

Keywords:
Length of stay
Regression models
Prediction
Machine learning
Clustering
Decision-making

ABSTRACT

Background: The General Medicine (GM) department has the highest patient volume and heterogeneity among other hospital specialties. Closely examining hospitalization data is crucial because patients come with various conditions or traits. Length of stay (LoS) in hospitals is often used as an efficiency indicator. It is influenced by various factors, including the patient's medical background, demographics, and type of diseases/signs/symptoms at the triage. LoS is a variable that can vary widely, making it difficult to estimate it promptly and accurately, but doing so is highly beneficial. Moreover, efficiently grouping and managing patients based on their expected LoS remains a significant challenge for healthcare organizations.

Objectives: This study aimed to compare the predictive ability of nine Machine Learning (ML) regression models in estimating the actual number of LoS days using demographics and clinical information recorded at admission as independent variables.

Methods: We analyzed data collected on patients hospitalized at the GM department of the Sant'Orsola-Malpighi University Hospital in Bologna, Italy, who were admitted through the Emergency Department. The data were collected from January 1, 2022, to October 26, 2022. Nine ML regression models were used to predict LoS by analyzing historical data and patient information. The models' performance was assessed through root mean squared prediction error (RMSPE) and mean absolute prediction error (MAPE). Moreover, we used K-means clustering to group patients' medical and organizational criticalities (such as diseases, signs, symptoms, and administrative problems) into four clusters. Feature Importance plots and SHAP (SHapley Additive exPlanations) values were employed to identify the more essential features and enhance the interpretability of the results.

Results: We analyzed the LoS of 3757 eligible patients, which showed an average of 13 days and a standard deviation of 11.8 days. We randomly divided patients into a training cohort of 2630 (70 %) and a test cohort of 1127 (30 %). The predictive performance of the different models was between 11.00 and 16.16 days for RMSPE and between 7.52 and 10.78 days for MAPE. The eXtreme Gradient Boosting Regression (XGBR) model had the lowest prediction error, both in terms of RMSPE (11.00 days) and MAE (7.52 days). Sex, arrival via own vehicle/walk-in, ambulance arrival, light blue risk category, age 70 or older, and orange risk category are some of the top features.

Conclusion: The ML models evaluated in this study reported good predictive performance, with the XGBR model exhibiting the lowest prediction error. This model holds the potential to aid physicians in administering appropriate clinical interventions for patients in the GM department. This model can also help healthcare services predict the resources necessary to better manage hospitalization.

* Corresponding author.
E-mail addresses: addisu.zeleke2@unibo.it (A.J. Zeleke), pierpaolo.palumbo@unibo.it (P. Palumbo), paolo.tubertini@aosp.bo.it (P. Tubertini), rossella.miglio@unibo.it (R. Miglio), lorenzo.chiari@unibo.it (L. Chiari).

https://doi.org/10.1016/j.imu.2024.101499
Received 30 January 2024; Received in revised form 28 March 2024; Accepted 12 April 2024
Available online 16 April 2024

1. Introduction

1.1. Importance of addressing hospital length of stay in general medicine patients

In General Medicine (GM), patients typically exhibit a wide range of conditions and varying characteristics, making it essential to examine hospitalization data closely. For this reason, predicting the Length of Stay (LoS) for each patient can be difficult due to the variability in their distribution. As hospitals face more pressure due to the global population aging, with growing demand and limited resources, predicting the LoS for patients at admission has gained popularity. A reliable prediction could aid, e.g., in efficiently scheduling surgeries and providing appropriate care for patients with prolonged LoS. Ensuring appropriate hospital LoS for patients in the GM department is crucial for many reasons. Firstly, longer hospital stays are linked to a higher risk of adverse events, such as hospital-acquired infections, falls, and pressure ulcers, which can negatively impact patient outcomes and well-being [1]. Extended hospitalization can also contribute to decreased physical function and a decline in overall health status. In addition to patient benefit, monitoring LoS is essential for patient satisfaction. Patients indeed prefer shorter hospital stays [2], associated with faster recovery and a reduced risk of iatrogenic complications. The use of resources is another critical factor, as more extended hospital stays require more nursing care, diagnostic tests, and medication administration. Extended stays can lead to increased costs for both the hospital and the patient and a higher burden on the healthcare system. However, in most cases, there is no knowledge about when GM patients will be discharged from the hospital. Accurate prediction of the LoS could hence enable hospitals to optimize healthcare resource management, resulting in better outcomes [3,4].

Patients are typically partitioned into subgroups or clusters based on their similar or shared characteristics, for instance, problems based on how long they were hospitalized. Due to the difficulties caused by heterogeneous patient populations, grouping them into homogeneous, understandable groups is helpful. Research has shown that this approach offers several benefits in improving hospital and health facility planning and management [5]. Additionally, partitioning or grouping patients based on their LoS can provide valuable insights into patient care and outcomes [2]. Patient grouping is an advantageous method to simplify and enhance our understanding of diverse patient populations [6]. Patient diagnosis is influenced by various factors such as illness severity, medical complications, recovery rate, hospital stay, resource consumption, discharge location, and social factors. However, it can be challenging to allocate resources effectively due to the heterogeneity among patients [7]. Several patient grouping methods have been developed to address this issue. They are reported in the literature, identifying homogeneous patient groups within a given hospital population [7–9]. In similar studies [10] concerning a machine-learning-based prediction of prolonged hospitalization in patients admitted to general settings that include all departments and specialties (referred to as the 'all-patients' model), the finding demonstrated that the XGBR model performed better. Furthermore, key features of significance or top feature importance were identified. In this study, we decided to take a closer look at GM-specific patients since this department has the highest patient volume and heterogeneity among other hospital specialties.

1.2. Related works

As part of clinical research, Machine Learning (ML) models generally aim to predict clinical outcomes based on multiple predictors. Patients can benefit from new and rapidly growing data sources by leveraging ML, Artificial Intelligence (AI), and other statistical methods. However, we had **little** cause to believe that any particular model class would be best for this problem. As such, in this paper, we systematically consider different ML regression models for predicting in-hospital LoS in GM patients.

Due to the wide range of ML regression studies conducted across various departments, specialties, diseases, and objectives, the selection of algorithms varies based on their respective performance disparities concerning data size. However, ensemble algorithms tend to outperform individual algorithms because they combine weak learning algorithms. Some similar works include the following. Siddiqa et al. [11] used several regression techniques, including multiple linear regression (MLR), decision tree regression (DTR), linear regression (LR), ridge regression (RR), eXtreme Gradient Boosting Regression (XGBR), and random forest regression (RFR) to predict LoS of inpatients. Their findings indicated that RFR was the most effective model, with an achieved mean square error (MSE) of 5 days. In another study involving 896 surgical patients, Chuang et al. [3] applied supervised ML models such as Local Gaussian Regression (LGR), support vector regression (SVR), and RFR to predict LoS. The findings indicated that the RFR model was the most accurate for predicting LoS. Kolchun et al. [12] conducted a separate study to develop a model for predicting passenger LoS following a motor vehicle accident. They evaluated several machine learning methods and found that a neural network (NN) algorithm had the lowest mean absolute error (MAE) of 2.23 days for predicting LoS. Similarly, Liu et al. [13] used data from seventeen hospitals in northern California and applied a mixture of regression models to predict LoS in hospitals. They demonstrated that Laboratory Acute Psychological Score (LAPS) and Comorbidity Point Score (COPS) helped improve the models' efficiency.

Our paper focuses specifically on regression for the predictive modeling of LoS, modeled as a continuous outcome. However, the question remains as to which methods are the most effective in refining the accuracy of LoS prediction models or understanding the LoS variability among patients with different entry diagnoses.

1.3. Aims

This study aimed to estimate and compare the predictive ability of different machine-learning regression models for predicting the LoS of patients in the GM department. By leveraging observable characteristics of patients, the model aimed to provide accurate estimations of how long a patient is likely to stay in the hospital. Different ML approaches produce mixed results, illustrating the need to carefully select a prediction model that effectively describes the observed data. Additionally, we had little reason to believe in advance that any particular class of model would be the most suitable for this type of study. As a result, in this article, we comprehensively considered nine ML regression models to predict the outcome of LoS. We also sought to better understand the relationship between hospital stay duration and the criticalities before and during hospitalization. Moreover, we aimed to use unsupervised learning techniques to cluster patients into well-defined subgroups based on the reported LoS with their medical criticalities or problems. These techniques can help uncover patterns and structures in the data, leading to valuable insights for patient care and treatment.

2. Materials and methods

2.1. Study design and population

We screened for eligibility all the admissions to the hospital through the emergency department (ED) of the public Sant'Orsola-Malpighi University Hospital in Bologna, Italy, between January 1, 2022, and October 26, 2022. For our analysis, we selected hospitalized GM department patients. The primary outcome of this study was hospital LoS. LoS was calculated as the number of days between admission and discharge. Approaching the prediction of LoS as a regression rather than a classification task can provide more accurate results and valuable insights for the targeted purpose. This allows for a continuous measure of LoS duration and avoids the limitations of forced categorization. This

actual continuous measure of LoS can better account for the underlying complexity and individual variability in hospital stays. It will also reduce the potential bias in the modeling process. Patients whose discharge information was missing or unknown in any case were excluded from the analysis. Fig. 1 illustrates the detailed flowchart for selecting patients and the reasons for excluding patients.

2.2. Independent variables

Based on the available information reported upon admission, the independent variables that can be used to predict a patient's LoS in the GM department include:

- Demographic characteristics (gender and age): which can be used to identify patterns and trends in the data;
- Mode of arrival/source of admission (includes Ambulance −118, own vehicle/walk-in, or any other means): the way a patient arrives at the hospital or the source of their admission can provide valuable insights into their condition and the potential LoS;
- Risk categories: The risk categories assigned to patients based on triage at the entrance (red, orange, light blue, green, white) can help identify patients who require more intensive care and may have a longer LoS; and
- Criticalities or problems: This includes both medical and administrative factors such as diseases, signs, symptoms, and administrative issues that can affect the patient's condition and LoS.

2.3. Clustering

In healthcare applications, efficiently grouping spells according to LoS remains challenging. This is primarily due to natural variability in the LoS distribution. LoS-driven patient grouping could significantly improve bed allocation planning and patient admission and discharge processes. It is still a research challenge to efficiently group patients based on their LoS. Numerous clustering techniques are available for grouping patients based on similarities, including density-based, k-means, hierarchical, and model-based clustering (such as the Gaussian Mixture model).

We searched clusters on patients' criticalities, using LoS mean and interquartile range as descriptive variables for each criticality. The k-means clustering algorithm randomly selected k initial criticalities as cluster centroids. Each criticality was then assigned to the nearest centroid, which was updated based on the mean of all the criticalities assigned to it. This iterative process was repeated until either the centroids no longer moved or a predetermined number of iterations were completed. By doing so, the algorithm seeks to find the best groupings of similar criticalities in terms of their LoS mean and IQR values.

2.4. Regression models

To predict the continuous outcome, we applied nine ML regression models. In-depth descriptions of these models are available in the literature [14]. However, we provide a concise overview of each model here.

1. *Linear regression (LR)* [15] predicts the outcome values using the feature values combined into a linear equation in the feature parameters. Feature parameters are chosen to maximize the likelihood of the observed outcome values or, equivalently, to minimize the residual sum of squares.
2. *Bayesian Ridge Regression (BRD)* is a linear Bayesian regression with a Gaussian prior distribution on the feature parameters. The standard deviation λ on the prior distribution works as a regularization parameter on the Euclidean norm of the parameter vector.

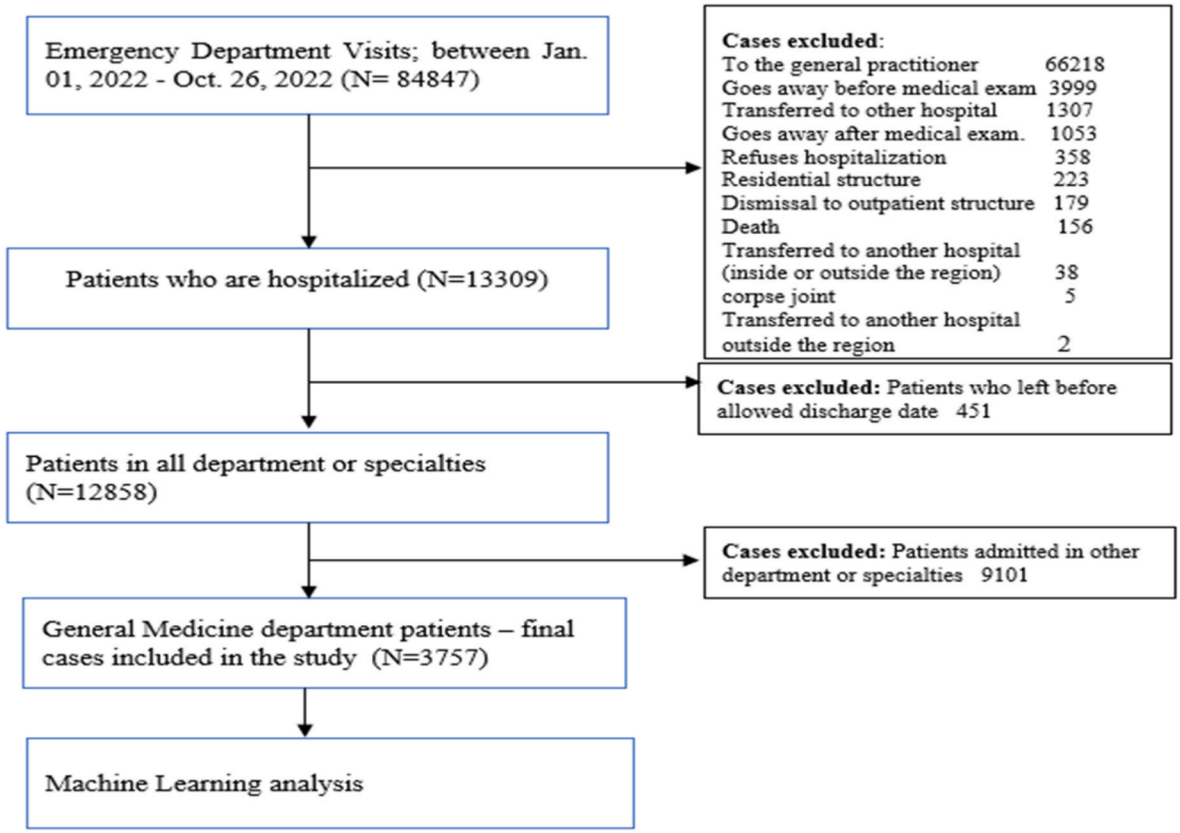


Fig. 1. Flowchart of patient selection.

3. *Decision tree regression (DTR)* [16] builds a tree-like arrangement of variables by dividing the data into smaller groups, resulting in a decision tree linked to these subsets of data.
4. *Random forest regression (RFR)* [17] is an ensemble learning technique that constructs multiple independent decision trees and aggregates their predictions. It is known to be robust to outliers and non-linear relationships in the data.
5. *Light Gradient Boosting Regression (GBR)* [18] is an ensemble learning technique that builds a model by iteratively combining weak learners to improve the accuracy of predictions. It is designed to provide fast and efficient training in decision tree models.
6. *Linear support vector regression (SVR)* [19] transforms data into a high-dimensional space using a kernel function and then finds a linear function that describes the relationship between input features and the target variable with a hyperplane.
7. *Extreme Gradient Boosting Regression (XGBR)* [20] is a DTR-based ensemble ML model (it combines the predictions from multiple machine learning algorithms to make more accurate predictions than any individual model) used to increase the speed and performance accuracy of the model.
8. *K-Nearest Neighbors (KNN)* [21] is an index-based algorithm that computes distances between instances, assigns indexes to these points, and stores the sorted distances and their indexes. It predicts the target value of a new instance as an average of the target values of k nearest training instances.
9. *Negative binomial regression* (NBR) [22] is a generalized linear model used for count outcomes, similar to Poisson regression. It incorporates an extra parameter to model over-dispersion, i.e., the count outcomes' variance is greater than their mean.

An overview of the regression prediction model framework is shown in Fig. 2.

2.4.1. Parameters and hyper-parameters tuning - grid search

In order to attain optimal performance, it is essential to determine the optimal hyper-parameters for each algorithm. A grid search was performed to identify the best set of hyper-parameters for each model. It tests all possible combinations in a parameter grid where one defines the possible values or ranges for each hyper-parameter. The grid search was performed with a 10-fold cross-validation nested within the training set. The optimal hyper-parameters discovered through the grid search for each algorithm are presented in Table 1.

2.4.2. Model evaluation

To assess the generalization and performance of the various models, we employed root mean squared prediction error (RMSPE) and mean absolute prediction error (MAPE), as indicated by equations (1) and (2) [23].

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

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

In these equations, y and $\hat{y}$ represent the observed and predicted LoS, respectively, and n the total number of patients in the test dataset. A single performance measure is usually insufficient to generalize a model's accuracy. This is because a single measure may not fully capture the complexity and diversity of the problem at hand, so considering additional measures is necessary to understand the model's performance better.

RMSPE is the square root of the average of the squared residuals, where a residual is the difference between a predicted value and an observed value. It indicates the standard deviation of residuals. On the other hand, MAPE is computed by taking the average of all the absolute residuals. RMSPE and MAPE are similar, although RMSPE places a higher penalty on large prediction errors than MAPE. The RMSPE metric is commonly employed in regression models as it can penalize larger errors while remaining easily interpretable.

2.4.3. Feature importance score and model interpretation

We calculated the importance of the features in the best-performing model using the gain score [14]. The gain score measures a feature's contribution to a model by evaluating how much it reduces the entropy or impurity of a split on that feature. This measure is commonly used for tree-based models. The higher the score, the more influential the feature is for making predictions. When a split decision is made with each feature, the gain score represents the average gain of each feature. We also used SHAP (Shapley Additive exPlanations) for model interpretation. SHAP is a Shapley-based novel approach that aids in explaining each feature's effect on model outcome [24]. In this explanation, the interplay among the local features is evaluated and subsequently aggregated to offer insights into individual predictors as well as the overarching model. Additionally, SHAP identifies predictors that degrade the model's performance and identifies high-risk and low-risk predictors. Finally, we analyzed the associations between the essential features and the impact of the LoS outcome using SHAP values to improve the interpretation of the learning results of the best-achieving model.

2.5. Tools

In this study, we employed Python version 3.8 and executed it on Jupyter Notebook. We leveraged the Scikit-learn (sklearn) library to carry out the analyses.

3. Results

3.1. Patient selection

Fig. 1 displays the flowchart outlining the patient selection process for inclusion in the analyses according to the study eligibility criteria.

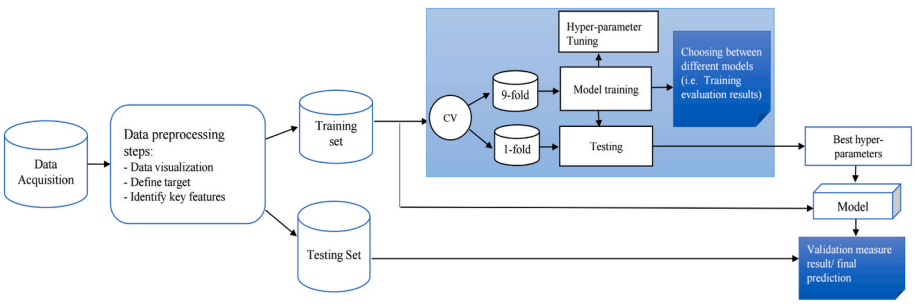


Fig. 2. The predictive framework based on regression models.

Table 1
Optimal Hyper-parameter combinations for ML Regression Models.

LR	DTR	RFR	Light GBR	Linear SVR	XGBR	KNN
–	Max depth: 10 Max features: Auto Criteria: Entropy Min sample split: 4 Min sample leaf: 2	n _estimators: 100 Max depth: 3 Max features: Auto Criteria: Entropy Min sample split: 4 Min sample leaf: 2	Learning rate: 0.1 Max depth: 5 n_estimators: 100	Kernel: linear Gamma: 0.1 C: 1	Criterion: MSE Learning rate: 0.1 Max depth: 5 n_ estimators: 100	n_ neighbors: 10

From January 1 to October 26, 2022, 84,847 patients were recorded across all departments in the ED. Upon applying the exclusion criteria, 12,858 patients were retained across all department and specialty settings. Ultimately, for the purpose of our analysis, we included 3757 patients admitted to the GM department.

3.2. Descriptive statistics

Of the 3757 patients analyzed, 50.4 % were males (49.6 % females), and 67.6 % were aged 70 or older. Their average LoS was 13 days. The entire cohort of patients was randomly split into two groups: a training cohort of 2630 patients (70 % of the total cohort) and a test cohort of 1127 patients (30 % of the total cohort). Table 2 presents a

Table 2
Descriptive statistics measures of predictors and LoS.

Variables	Total (n = 3757)	%	Length of Stay (in days)					
			Mean	Med	Std	Min	IQR	Max
Age, years:								
0 to 17	3	0.08	7.66	6	5.69	3	4.5–10	14
18 to 29	53	1.41	11.75	8	13.24	0	5–12	80
30 to 49	284	7.56	11.84	9	9.68	0	6–15	85
50 to 69	876	23.32	13.49	10	12.43	0	6–17	104
70 and older	2541	67.63	13.36	10	11.80	0	6–16	143
Gender:								
Male	1893	50.39	12.75	10	11.02	0	6–16	143
Female	1864	49.61	13.76	10	12.59	0	6–17	119
Mode of arrival:								
Ambulance	2225	59.22	12.83	10	11.32	0	6–16	143
Own vehicle/walk-in	1049	27.92	13.79	10	12.74	0	6–17	114
Others	483	12.86	14	10	12.74	0	7–17	119
Triage Category:								
Red	178	4.74	11.97	9	11.56	0	5–16	99
Orange	969	25.79	12.89	10	10.85	0	6–16	88
Light blue	1634	43.49	13.79	10	12.66	0	6–17	143
Green	936	24.91	13.01	10	11.41	0	6–16	114
White	40	1.06	10.80	9	9.21	0	4–12.3	51
Criticalities/Problems:								
Diseases/Signs/Symptoms								
Dyspnea	703	18.71	12.66	9.5	6.20	2	6–16	99
Fever/hyperpyrexia/hyperthermia	406	10.81	13.83	10	11.68	0	7–19	85
Abdominal pain	390	10.38	12.97	10	11.7	0	6–16	90
Non-specific minor disorders	201	5.35	14.21	11	12.86	0	6–16	80
Generalized asthenia	155	4.13	14.90	11	16.24	0	7–16	143
Nausea or vomiting repeated	95	2.53	13.37	10	13.17	0	5–17.5	88
Syncope/pre-syncope	115	3.06	11.29	9	9.67	2	6.5–14	83
State of confusion	103	2.74	11.41	10	9.2	0	6–14	66
Altered level of consciousness	89	2.37	13.29	11	11.21	0	7–16	78
Pallor/anemia	70	1.86	10.87	9	7.1	0	6–13	33
Hematochezia/rectal bleeding/melena	82	2.18	11.13	8.5	9.24	2	5.3–13	55
CVD associated with any symptoms	79	2.10	11.34	9	7.03	1	6–15	30
Chest pain of suspected CV cause	77	2.05	10.44	8	7.26	2	5–13	43
Swollen/edematous leg	65	1.73	16.40	12	13.5	4	7–20	64
Diarrhea	58	1.54	11.08	9.5	6.20	2	7–14	29
Heart palm/irregular wrist	48	1.28	10.71	7	10.5	1	5–12	58
Polytraumas - contusive	48	1.28	14.25	12	10.61	1	8–17.3	47
Cough/congestion	47	1.25	14.31	10	11.07	0	6.5–21	46
Diagnostics biochemical examinations	45	1.20	12.71	11	8.92	0	6–17	51
Urinary tract infection symptoms	42	1.12	14.11	12.5	8.29	2	8–18	35
Pain at the side	39	1.04	9.35	9	5.29	1	5.5–12	25
Chest pain not suspected by CV cause	30	0.80	11.26	9.5	8.34	3	6.3–12	42
Macro-hematuria	28	0.75	14.53	10.5	12.00	2	6.8–17	49
Lower limbs pain	24	0.64	15.87	14.5	7.50	4	10–23	29
Head trauma	19	0.51	11.47	8	15.04	2	6–10.5	71
Lower limbs injury	7	0.19	19.57	16	8.82	14	15–19	39
Others	594	15.81	14.43	11	13.18	0	6–19	119
Administration problems								
Request for urgent specialist advice	48	1.28	14.50	10.5	11.74	2	7–17.3	56
Request for prescription or performance	50	1.33	17.48	11	22.33	2	6.3–21	117

comprehensive summary of the descriptive statistics.

The LoS ranged from 0 to 143 days, zero days meaning that admission and discharge occurred during the same day. The LoS values show a right-skewed distribution, with most values ranging from about 1 to 18 days (as shown in Fig. 3, left panel). Lower limb injury, prescription/performance request, swollen/edematous leg, lower limb pain, generalized asthenia, macro-hematuria, and request for urgent specialist advice were the top problems with the longest average LoS (see Fig. 3, right panel**)**. The count plot of each predictor is shown in Appendix S1. Additionally, a granular analysis of the LoS was performed by type of disease/signs/symptoms or administration issues. Fig. 4 presents the boxplot of hospital LoS for each criticality or problem. Unexpected records, i.e., outliers, were removed from the graph for effective comparative analysis.

3.3. Clustering

Fig. 5 depicts the result of clustering groups based on the mean and IQR of the LoS for each cluster. The plot shows the mean LoS on the horizontal axis and the IQR on the vertical axis, with k-means clustering applied to a dataset containing these values. Choosing the right number of clusters is a crucial step in k-means clustering since it impacts the quality and interpretability of the results. The Elbow Method, Silhouette Score, Silhouette Analysis, Domain Knowledge, and many other techniques can all be used to determine the number of clusters. However, we decided to select four clusters based on the data set and using domain knowledge. Distinct colors represent patient criticalities assigned to different clusters.

Table 3. Illustrates the parameters obtained from the four-cluster analysis. A large portion of the patients' criticalities (37.9 %) were allocated to clusters 1 and 3, while only one criticality was assigned to cluster 4.

3.4. Regression models

The evaluation metrics for the ML regression models are presented in Table 4. The XGBR models exhibited superior performance on the test set, displaying the lowest prediction error with an RMSPE of 11.00 and MAPE of 7.52 among the various methods examined. These results highlight the effectiveness of XGBR in accurately predicting the LoS. While XGBR models have proven effective in solving a range of machine learning problems, it is crucial to perform careful hyperparameter tuning to attain optimal performance. On the other hand, the DTR model demonstrated the smallest prediction error in the training set but the highest RMSPE and MAPE values in the test set, suggesting poor generalization performance and overfitting to the training data, making it unsuitable for this specific problem.

In Fig. 6, the RMSPE and MAPE of the nine models are visualized in a single plot to ease comparison. The test data show that the XGBR model performs better, suggesting it may be more reliable and generalizable to new patients or data.

A calibration curve cannot be used since ML regression models do not provide direct probability estimates. However, we can still assess the calibration of predicted values in a regression model using a different approach. One common technique is to group the predicted values into bins and calculate the mean predicted value and the corresponding mean actual value for each bin. We can then plot these values to visualize the calibration. Fig. 7 shows that the optimal model (i.e., XGBR) displayed a tendency to underestimate hospital stays for cases with short LoS predictions while overestimating hospital stays for cases with long LoS predictions. The calibration plots in the test set are presented in Figure S3 of the Appendix for the other models.

The XGBR model, selected as the best model, has feature importance scores shown in Fig. 8. These scores quantify the extent to which a feature contributes to the overall reduction of the objective function, such as mean squared error, in a tree-based model. The more points a feature receives, the more significant it is when making predictions. The gain score, which shows the average gain of each feature when utilized in a split decision, is one of the feature importance ratings in the model. Gain results from the feature in the tree's ability to reduce the goal function. As shown in Figure S2 in the Appendix, the heatmap plot of the similarity matrix shows the absence of collinearity between the top important features. A darker shade indicates greater similarity between the features, whereas a lighter shade indicates a more minor similarity.

SHapley Additive exPlanations (SHAP) were applied to the final XGBR model to explain the effect of each feature variable on the target variable LoS. TreeExplainer was used to generate the SHAP values. A beeswarm plot was created to depict the significance of the global characteristics, as shown in Fig. 9. This plot shows the overall effects of each feature on the model's output. The x-axis shows the impact of a feature on prediction, while the color shows its value. A red feature has a high value and is positively correlated with the prediction, whereas a blue feature has a low value and is negatively correlated with the prediction. SHAP value plots show how each feature contributed to the model's output and can help identify the most relevant features and their contributions to specific predictions.

Both the importance score analysis and SHAP list sex, arrival to ED by ambulance, and age equal to or greater than 70 among the top five

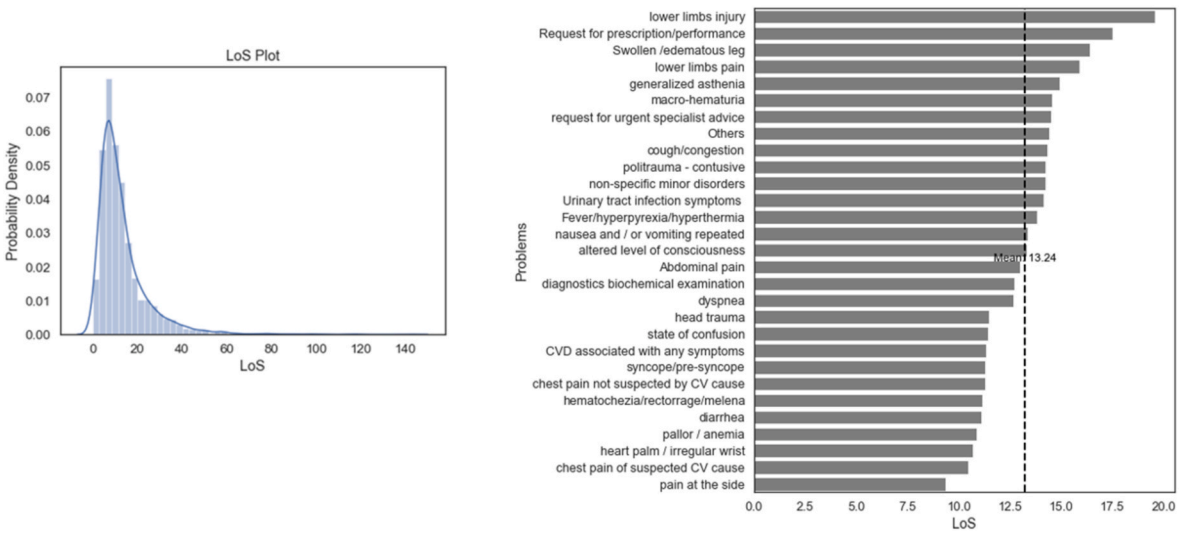


Fig. 3. The LoS distribution plot (left panel) and the average LoS for each patient issue (right panel). CVD: cerebral vascular disease; CV: cardiovascular.

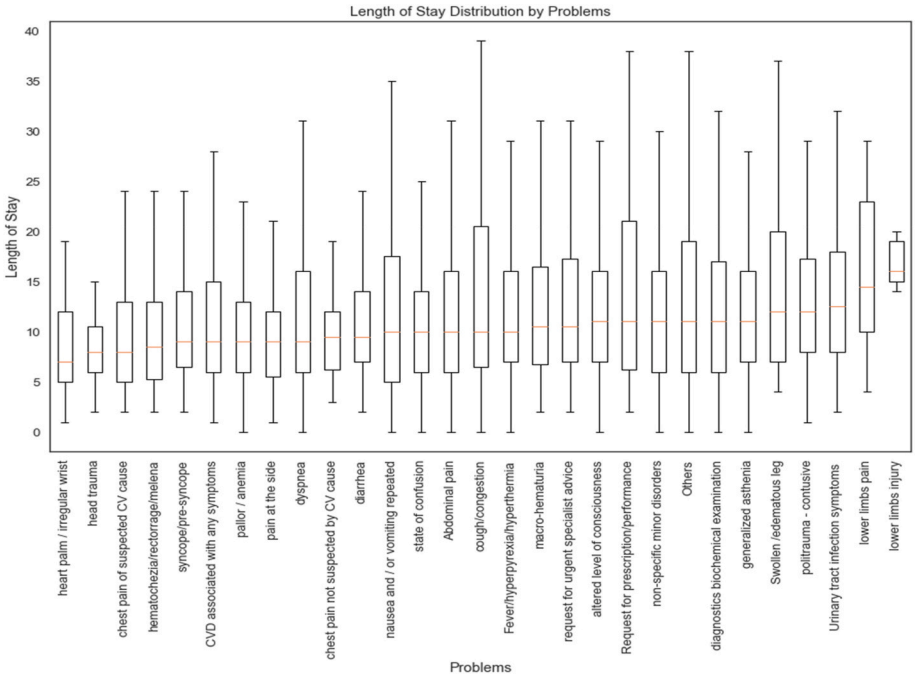


Fig. 4. A granular analysis of the LoS by criticalities or problems.

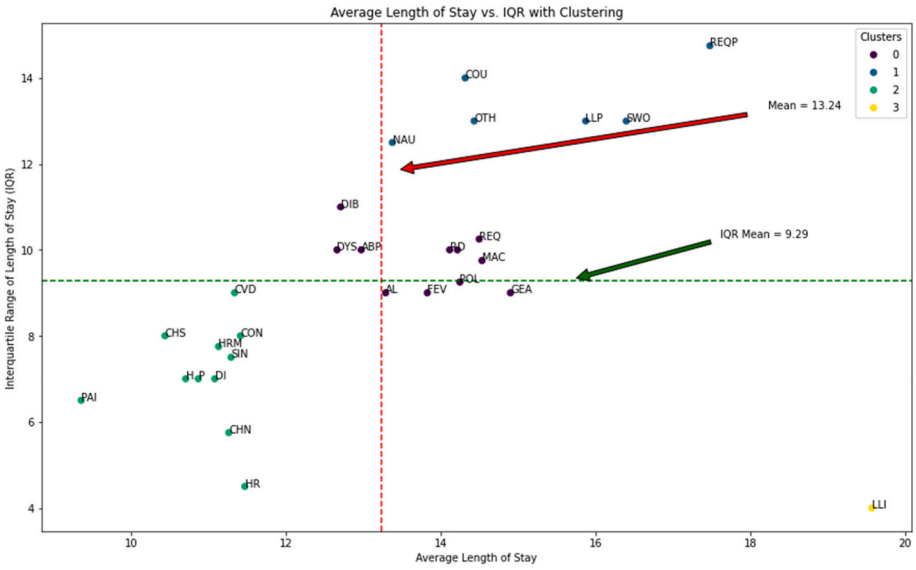


Fig. 5. Clustering on patients' criticalities according to the average and IQR of LoS. CVD; CVD associated with any symptoms, R; urinary tract infection symptoms, SWO; swollen/edematous leg, CON; state of confusion, SIN; syncope/pre-syncope, REQ; request for urgent specialist advice, REQP; Request for prescription/performance, POL; politrauma – contusive, P; pallor/anemia, PAI; pain at the side, OTH; Others, D; non-specific minor disorders, NAU; nausea and/or vomiting repeated, MAC; macro-hematuria, LLP; lower limbs pain, LLI; lower limbs injury, HRM; hematochezia/rectorrage/melena, H; heart palm/irregular wrist, HR; head trauma, GEA; generalized asthenia, FEV; fever/hyperpyrexia/hyperthermia, DYS; dyspnea, DI; *diarrhea,* DIB; diagnostics biochemical examination, COU; cough/congestion, CHS; chest pain of suspected CV cause, CHN; chest pain not suspected by CV cause, AL; *altered* level of consciousness, ABP; abdominal pain.

Table 3
Cluster-level parameter measures (NaN–No available SD LoS for a single criticality).

Clusters	Percentage	Average LoS	Average IQR	SD LoS	SD IQR
Cluster 1:	37.9	10.95	7.09	0.61	1.22
Cluster 2:	20.7	15.31	13.38	1.53	0.83
Cluster 3:	37.9	13.82	13.82	0.78	0.63
Cluster 4:	3.5	19.57	4.00	NaN	NaN

most influential variables.

4. Discussion

This study aimed to develop a machine-learning regression model capable of predicting the length of stay (LoS) for patients in the GM department. The results of the present study showed that the XGBR model performed better than the other eight ML regression models at predicting the actual number of LoS for patients with GM-specific problems. The necessity of carefully picking a prediction model that

Table 4
Comparisons of the ML regression models (i.e., evaluation metrics).

Models	Training Set		Test Set	
	RMSPE	MAPE	RMSPE	MAPE
LR	11.94	7.94	11.02	7.56
Bayesian Ridge	12.03	7.98	11.04	7.53
DTR	5.74	2.06	16.16	10.78
RFR	6.94	4.23	12.38	8.48
Light GBR	10.72	7.39	11.77	7.59
Linear SVR	11.01	7.40	11.10	7.60
XGBR	11.64	7.68	**11.00**	**7.52**
KNN	10.91	7.35	12.06	8.43
NBR	11.94	7.93	11.01	7.54

effectively describes the observed data is demonstrated by different ML approaches, giving mixed findings. Moreover, we also had little reason to believe in advance that any particular class of model would be the best choice for this type of study. As a result, in this article, we have comprehensively considered nine ML regression models to predict the LoS outcome. Furthermore, the primary objective of this study was to estimate and compare the predictive ability of different machine-learning regression models for predicting the LoS of patients in the GM department. Accurate LoS prediction models could facilitate the proactive allocation of vital healthcare resources and have the potential to reduce unwarranted hospital stays, with a particular focus on AOSP Bologna.

In a previous study [11], we analyzed the 'all-patients' model, including patients admitted in all departments; however, in this study, we focused on GM-specific patients, who have the highest volume and heterogeneity among other specialties in the hospital. Since patients in this department have various diseases or characteristics, careful analysis of hospitalization data is essential. Although it can be challenging to estimate LoS precisely, doing so is quite advantageous. Nevertheless, healthcare applications still face a considerable barrier to properly organizing patient spells depending on their LoS. A LoS is a complex variable influenced by several clinical and social factors [25]. Modeling is one part of healthcare responses. Nowadays, healthcare systems have started focusing on efficient resource management and the prediction of feature outcomes. This is to ensure optimal levels of care, lower associated costs, and enhance patient care [26].

Before assessing the models' performance, we searched for the best hyper-parameters for each model (i.e., a grid tuning hyper-parameter), considering RMSPE and MAPE as loss function measures. The XGBR model outperformed the other eight models, with RMSPE of 11 days and MAPE of 7.52 days. This suggests that XGBR may be more reliable and generalizable to new data than other regression models. As we searched for similar studies, ensemble regression models like XGBR, light GBM, and RFR usually produced superior results than single constituent models [27]. Another study [10] used six ML regression models, such as MLR, DTR, LR, RR, XGBR, and RFR, to predict LoS in a hospital. The study concluded that the RFR model outperformed the other models and achieved the lowest MSE of five days. A LoS study for patients with sepsis [28] compared six ML regression models, namely LR, RFR, KNN, NN, XGBR, and light GBR, and found that light GBM showed the best

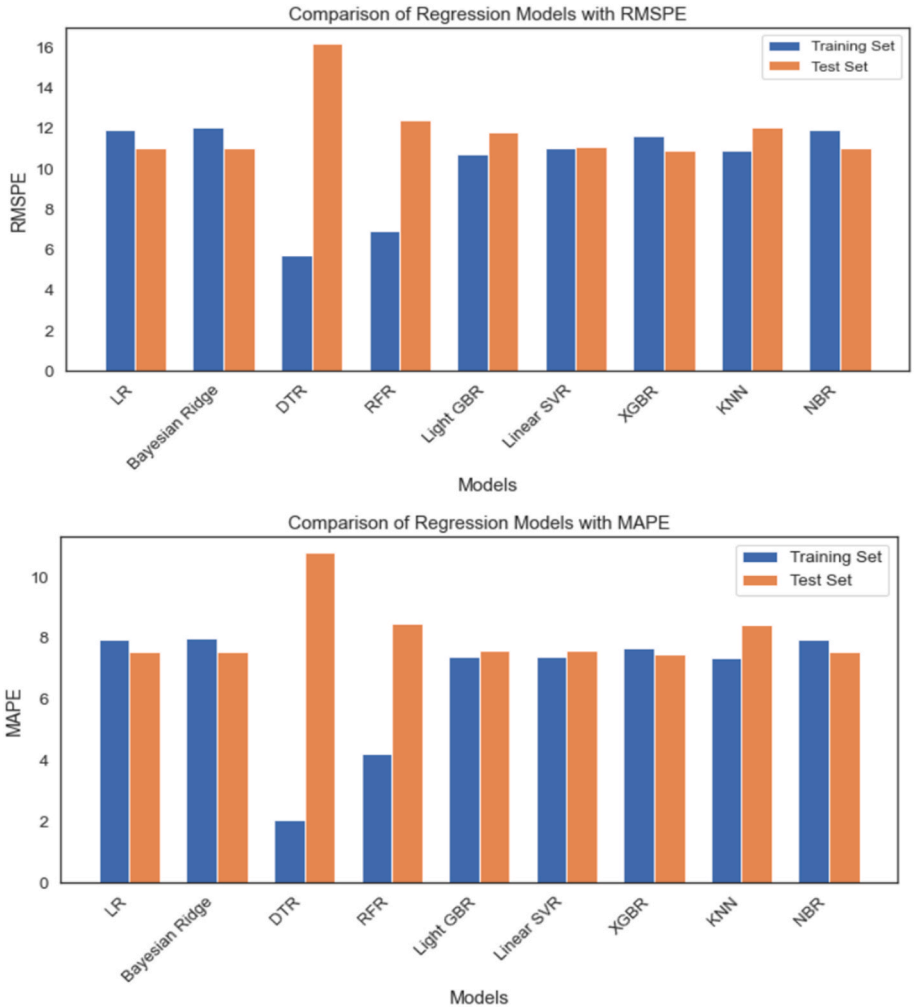


Fig. 6. Comparative analysis of regression models, with RMSPE displayed on the top panel and MAPE on the bottom panel.

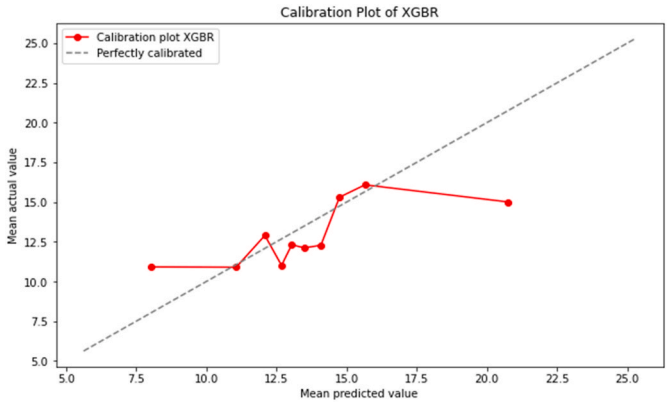


Fig. 7. Calibration plots for best ML regression model.

model having an MAE of 4 days. Similarly, in another study [29], light GBR showed a lower prediction error MAE of 3 days in inpatients.

There are several advantages of XGBR over other ML methods, such as the fact that it does not assume data distribution, uses decision trees, and can, therefore, be unaffected by multicollinearity. Ensemble methods, like XGBR, also automatically estimate feature importance from a trained predictive model, resulting in a boosted decision tree score for each feature. Consequently, the most important features are identified based on the chosen model, which includes sex, arrival mode (by own means/walk-in or by ambulance), triage category (light blue, orange, green, or red), age group (70 and older, 50 to 69, or 30 to 49), and specific conditions such as hematochezia/rectal bleeding/melena, pain at the side, pre-syncope, pallor/anemia, generalized asthenia, and prescriptions/performance requests. However, the values of performance measures can vary depending on the dataset and context. It is, therefore, essential to conduct rigorous testing and validation to ensure that the chosen model is appropriate for the task at hand. Additionally, other studies may use different sets of performance measures or evaluate different subsets of regression models, making direct comparisons difficult.

The findings of the present study have the potential to assist in predicting the future LoS for new patients. In this way, healthcare providers or hospital administrators can estimate and manage their resources more effectively. Patients can also benefit from knowing how long they will be staying in the hospital, estimating their treatment costs, and managing their other personal affairs in a timely manner. Also, the findings can help other researchers assess the performance of different regression models. The other benefit of this study is that GM is the most prevalent and heterogeneous department, which makes it essential for a hospital to consider when deciding how long to keep new patients in the facility.

4.1. *Limitations of the study*

It is important to note that this study has several limitations. This study was limited by the fact that a single cohort was used to develop and test the model, meaning that no external validation was possible. Despite promising results, external validation is vital to ensuring the model's generalization abilities [30]. In other words, evaluating the model's performance on different datasets and populations is necessary to ensure its robustness and reliability. As another limitation, the study did not include vital signs or laboratory test results for triage assessment, which could be crucial for better predicting hospital stay length.

5. Conclusions

XGBR was the best-performing model out of the nine ML regression models, with the lowest RMSPE and MAPE. The proposed method performs considerably in predicting LoS in the GM department, which is expected to provide valuable support for clinical decision-making. The most influencing variables for LoS prediction were sex, mode of arrival to the ED, and old age. Patients' criticalities were clustered into four groups according to their LoS average and interquartile range.

LoS prediction models could be highly advantageous to healthcare providers, as they can leverage them to accurately predict the duration of a patient's hospital stay, ensuring top-notch care and sufficient resources for all patients.

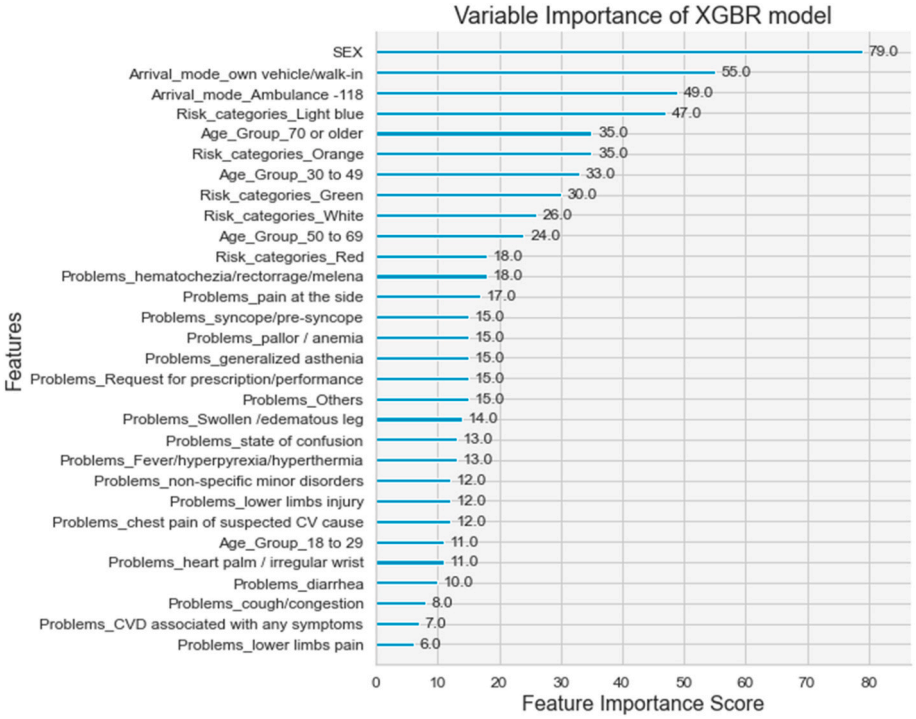


Fig. 8. Feature importance score (CVD: cerebral vascular disease; CV: cardiovascular).

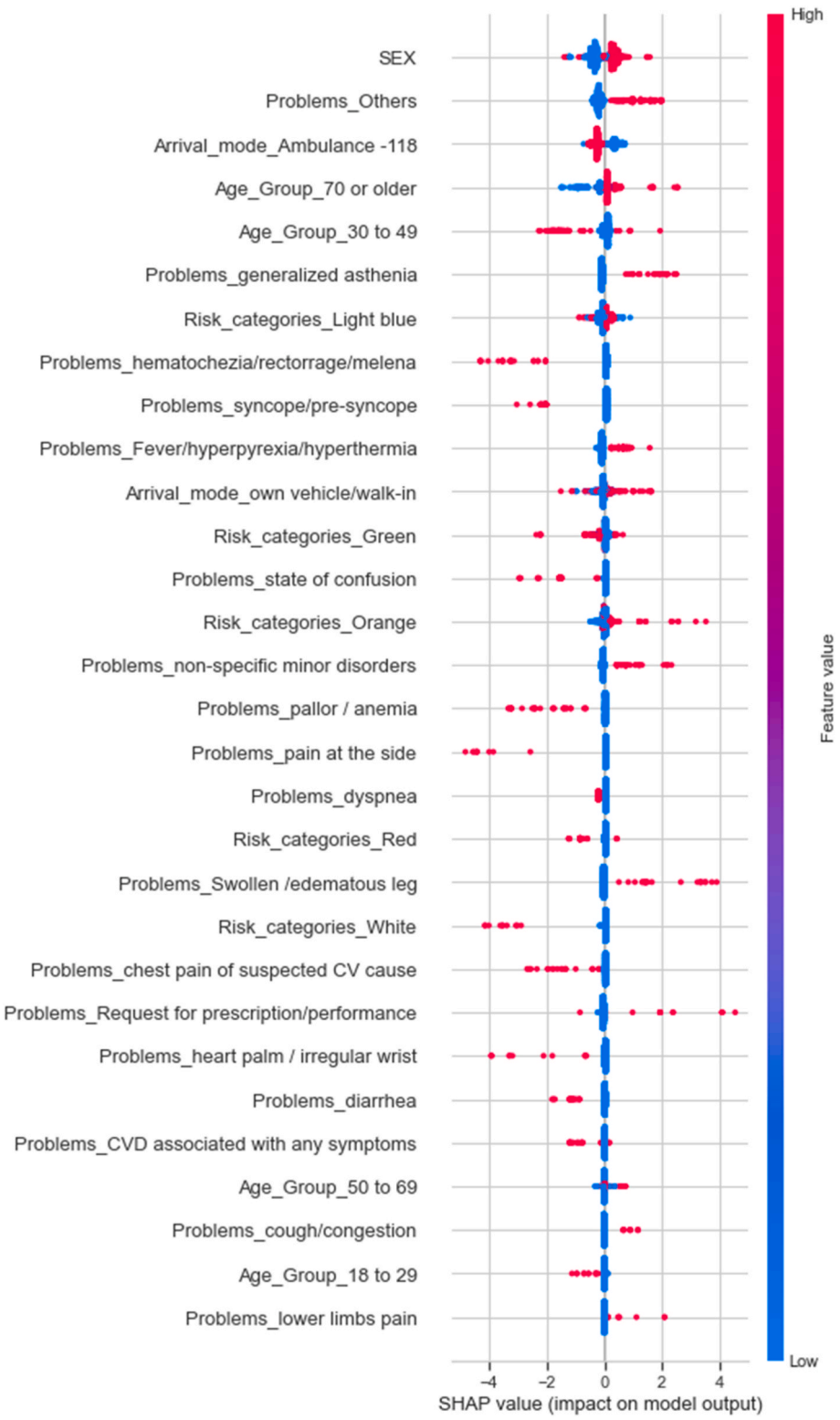


Fig. 9. Beeswarm plots based on the SHAP values for model interpretations (CVD: cerebral vascular disease, CV: cardiovascular).

Ethical considerations

The University of Bologna's (Italy) bioethics Committee reviewed and granted ethical approval for the studies involving human participants (approval number 0058022, February 24, 2023).

Funding

This research was partly supported by Policlinico Sant'Orsola-Malpighi through funding of the Ph.D. scholarship of AJZ.

Informed consent statement

Not applicable.

Data availability statement

The data are not publicly available due to ethical restrictions.

CRediT authorship contribution statement

Addisu Jember Zeleke: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Pierpaolo Palumbo:** Writing – review & editing, Software, Formal analysis, Data curation. **Paolo Tubertini:** Writing – review & editing, Resources, Project administration, Investigation, Data curation. **Rossella Miglio:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Lorenzo Chiari:** Writing – review & editing, Supervision, Project administration, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

Our appreciation goes out to all the personnel involved in data organization activities at Policlinico Sant'Orsola-Malpighi, which encompasses nurses, doctors, and other administrative staff.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2024.101499.

References

[1] Zolbanin HM, Davazdahemami B, Delen D, Zadeh AH. Data analytics for the sustainable use of resources in hospitals: predicting the length of stay for patients with chronic diseases. Inf Manag 2022;59(5):103282. https://doi.org/10.1016/j.im.2020.103282.
[2] Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S. Analysis of length of hospital stay using electronic health records: a statistical and data mining approach. PLoS One 2018;13(4):e0195901. https://doi.org/10.1371/journal.pone.0195901.
[3] Chuang M, Hu Y, Lo C. Predicting the prolonged length of stay of general surgery patients: a supervised learning approach. Int Trans Oper Res 2018;25(1):75–90. https://doi.org/10.1111/itor.12298.
[4] Lamere AT, Nguyen S, Niu G, Olinsky A, Quinn J. Predicting the Length of Stay in Hospital Emergency Rooms in Rhode Island 2021:35–48. https://doi.org/10.1108/S1477-407020210000014004.
[5] Sanderson H. The development of patient groupings for more effective management of health care. Eur J Publ Health 1997;7(2):210–4. https://doi.org/10.1093/eurpub/7.2.210.
[6] Harper PR. A review and comparison of classification algorithms for medical decision making. Health Pol 2005;71(3):315–31. https://doi.org/10.1016/j.healthpol.2004.05.002.
[7] Harper PR. A framework for operational modelling of hospital resources. Health Care Manag Sci 2002;5(3):165–73. https://doi.org/10.1023/A:1019767900627.
[8] Churilov L, Bagirov A, Schwartz D, Smith K, Dally M. Data mining with combined use of optimization techniques and self-organizing maps for improving risk grouping rules: application to prostate cancer patients. J Manag Inf Syst 2005;21(4):85–100. https://doi.org/10.1080/07421222.2005.11045826.
[9] Costa AX, Ridley SA, Shahani AK, Harper PR, de Senna V, Nielsen MS. Mathematical modelling and simulation for planning critical care capacity. Anaesthesia 2003;58(4):320–7. https://doi.org/10.1046/j.1365-2044.2003.03042.x.
[10] Zeleke AJ, Palumbo P, Tubertini P, Miglio R, Chiari L. Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis. Front Artific Intellig 2023;6. https://doi.org/10.3389/frai.2023.1179226.
[11] Siddiqa A, Abbas Zilqurnain Naqvi S, Ahsan M, Ditta A, Alquhayz H, Khan MA, Adnan Khan M. Robust length of stay prediction model for indoor patients. Comput Mater Continua (CMC) 2022;70(3):5519–36. https://doi.org/10.32604/cmc.2022.021666.
[12] Kolcun JPG, Covello B, Gernsback JE, Cajigas I, Jagid JR. Machine learning to predict passenger mortality and hospital length of stay following motor vehicle collision. Neurosurg Focus 2022;52(4):E12. https://doi.org/10.3171/2022.1.FOCUS21739.
[13] Liu V, Kipnis P, Gould MK, Escobar GJ. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. Med Care 2010;48(8):739–44. https://doi.org/10.1097/MLR.0b013e3181e359f3.
[14] Hastie T, Tibshirani, R. Friedman, J. (n.d.). The elements of statistical learning. Data mining, inference, and prediction..
[15] Freedman DA. Statistical models. Cambridge University Press; 2009. https://doi.org/10.1017/CBO9780511815867.
[16] Loh W. Classification and regression tree methods. In: Wiley StatsRef: statistics reference online. Wiley; 2014. https://doi.org/10.1002/9781118445112.stat03886.
[17] Tin Kam Ho. (n.d.). Random decision forests. Proc 3rd Int Conf Document Analys Recognit, 278–282. https://doi.org/10.1109/ICDAR.1995.598994.
[18] Ke, G. , Meng, Q. , Finley, T. , Wang, T. , Chen, W. , Ma, W. , Ye, Q. , & Liu, T. (2017). (n.d.). LightGBM: a highly efficient gradient boosting decision tree. Neural Inform Proc Syst..
[19] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2(2):121–67. https://doi.org/10.1023/A:1009715923555.
[20] Chen T, Guestrin C. XGBoost. Proc 22nd ACM SIGKDD Int Conf Knowledge Discov Data Min 2016:785–94. https://doi.org/10.1145/2939672.2939785.
[21] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Statistician 1992;46(3):175. https://doi.org/10.2307/2685209.
[22] Cameron AC, Trivedi PK. Regression analysis of count data. Cambridge University Press; 2013. https://doi.org/10.1017/CBO9781139013567.
[23] Duan N. Smearing estimate: a nonparametric retransformation method. J Am Stat Assoc 1983;78(383):605–10. https://doi.org/10.1080/01621459.1983.10478017.
[24] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.
[25] Shea S, Sideli Rv, DuMouchel W, Pulver G, Arons RR, Clayton PD. Computer-generated informational messages directed to physicians: effect on length of hospital stay. J Am Med Inf Assoc 1995;2(1):58–64. https://doi.org/10.1136/jamia.1995.95202549.
[26] Garg L, McClean SI, Barton M, Meenan BJ, Fullerton K. Intelligent patient management and resource planning for complex, heterogeneous, and stochastic healthcare systems. IEEE Trans Syst Man Cybern Syst Hum 2012;42(6):1332–45. https://doi.org/10.1109/TSMCA.2012.2210211.
[27] Gabriel RA, Harjai B, Simpson S, Du AL, Tully JL, George O, Waterman R. An ensemble learning approach to improving prediction of case duration for spine surgery: algorithm development and validation. JMIR Perioperat Med 2023;6:e39650. https://doi.org/10.2196/39650.
[28] Chen L, Klasky HB. Six machine-learning methods for predicting hospital-stay duration for patients with sepsis: a comparative study. SoutheastCon 2022;2022:302–9. https://doi.org/10.1109/SoutheastCon48659.2022.9764052.
[29] Zeng X. Length of stay prediction model of indoor patients based on light gradient boosting machine. Comput Intell Neurosci 2022;2022:1–14. https://doi.org/10.1155/2022/9517029.
[30] Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;i3140. https://doi.org/10.1136/bmj.i3140.