Predicting the length of stay of patients in hospital applying machine learning techniques

Word count (3907)

1. Abstract

This project studies the applicability of machine learning techniques to predict the length of stay (LOS) of patients in the hospital. The Random Forest algorithm was used to predict LOS using data from the MIMIC III database that contains recordings of patients admitted to the Beth Israel Deaconess Medical Center in the critical units. The aim of this study was to identify the main determinants of the length of stay of patients in the hospital and check the effectiveness of the Random Forest algorithm to predict LOS. Data preprocessing and feature selection methods were used, and the model was evaluated with different performance metrics. The results stated that increasing the number of features in the Random Forest model enhances its performance to predict the length of stay of patients in the hospital. The number of comorbidities in patients was found as the most important feature to predict LOS. A positive correlation between predicted and actual length of stay was observed in the visualisation of predicted results. On the other hand, there were some outliers that deviated from the overall pattern, potentially signalling limitations of the models. Despite the restrictions of the study, the results could be used to improve patient treatment and maximise the efficient use of hospital resources.

2. Introduction

In most developed countries population ageing, implies persistent requests for hospital services. Health institutions and government authorities make every effort to adjust benefits to achieve more desirable results of patients with execution of healthcare procedures in accordance with standards. Evidence of effectiveness in care and performance of hospitals is determined by the length of stay (LOS) in hospital. Generally, less resource consumption and inexpensive cost per discharge is a consequence of patient short stay (Tsai et al., 2016). Bedside staffing required for patient care consumes most of the costs related to patients in the Intensive Care Unit (Straney et al., 2017). Predicting LOS has a massive impact in economic terms, but it is not limited to this. Its advantages extend to various aspects of care, patients and environmental impacts (Zolbanin et al., 2022).

Daily, the healthcare industry generates a large amount of data that includes a range of sectors and knowledge. These data contain records from disease diagnosis, medical condition of patients to healthcare resources. Therefore, different formats and types are present in the data, the above characteristics can impact the quality in the production of hospital data due to deficient practices and inconsistency (Neto et al., 2020).

These scenarios require a Machine Learning approach that automates LOS, enabling it to identify patterns and correlations containing interesting results that can be helpful in the decision making with challenge data related to the healthcare sector (Suha and Sanam, 2022). Machine Learning's unique ability for nonlinear fitting and superior predictive characteristics are the reasons for its broad application to forecast LOS. A bagged regression trees model developed by Xie et al. to predict LOS with data from insurance claims, found that demographic information contributed less to the LOS than medical record data. Nowadays, use of machine learning models to predict LOS have increased but the vast majority of studies applying these models cover patients with particular illness (e.g. cardiovascular disease) which limit its applicability (Hu et al., 2022).

In this study, the Machine Learning algorithm used is Random Forest (RF) to predict length of stay of patients in hospital. It was applied to data from the MIMIC (Medical Information Mart for Intensive Care) III database which incorporates recordings of patients admitted to Beth Israel Deaconess Medical Center in the critical care units. The aim of this study is to Identify the main determinants of LOS and verify Random Forest effectiveness to predict LOS.

3. Methodology

3.1 Data Collection

For this study the data was collected from MIMIC III database (a large and freely-available relational database) which stores recordings of patients (over 40 000) admitted to the Beth Israel Deaconess Medical Center in its critical care units from 2001 to 2012. The database data are de-identified health-related and it contains information such as mortality (both in and out of hospital), procedures, demographics, vital sign measurements and laboratory test results. In order to manage the data with adequate respect and care, considering that it includes patients' clinical care information there were requirements to meet, prior to access MIMIC III; become credential user of PhysioNet and complete a training course about human subject research (The Medical Information Mart for Intensive Care, 2023).

3.2 Data Preprocessing

MIMIC III consists of 26 tables, divided in four groups; dictionaries, define and/or track patient stays, data collected in the critical care unit and data selected in the hospital record system. To achieve the aim of this study, tables were selected from the hospital record system and track patient stays categories. These tables were connected by HADM_ID (identifier) which describes unique hospital admission.

The initial data analysis allowed to identify 58976 rows (patients) and 27 columns (variables), that included 17 numerical and 10 categorical data about hospital recording on patients. Columns with missing values were found and filled, numerical data filled with next or previous values while in the categorical data replaced with most common values.

Length of stay of patients were calculated with two variables, admission time and discharge time:

$$length\ of\ stay\ =\ discharge\ time\ -\ admission\ time$$

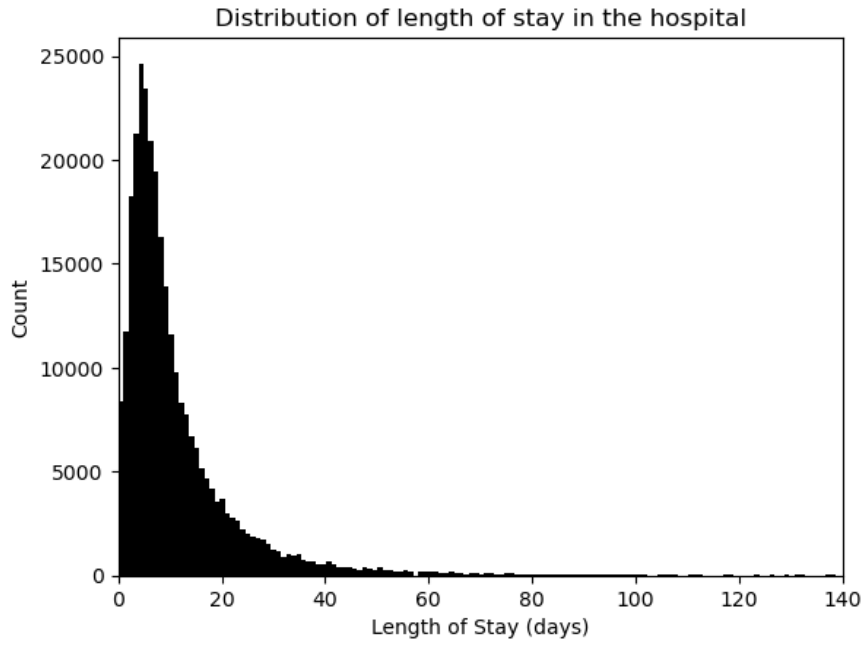

Figure 1. Distribution of length of stay in the Hospital

Patients age were determined by date of birth and admission time:

$$Age\ =\ admission\ time\ -\ date\ of\ birth$$

Age of patients ranged from 0 to 89, it was categorised into four age groups, children (0 - 15), young adults (16 - 30), adults (31 - 60) and seniors (61 - 90).

To predict the length of stay (days) of patients in the hospital, many factors or variables are frequently considered. These variables can involve demographic information about the patients such as ethnicity, marital status, religion and discharge location. In the dataset, these variables were segmented into more comprehensive groups. For example, marital status to married and unmarried, discharge location to home, home health care, hospital and other. This recategorization allowed a more nuanced understanding of how these features impact the length of stay for patients in the hospital.

The Diagnoses table contained an ICD-9 codes' column, with 6984 unique entries; these codes were classified into 17 disease and related health problem categories corresponding to each code. This new column (disease classification) was used with hospital admission to estimate the number of comorbidities per patient.

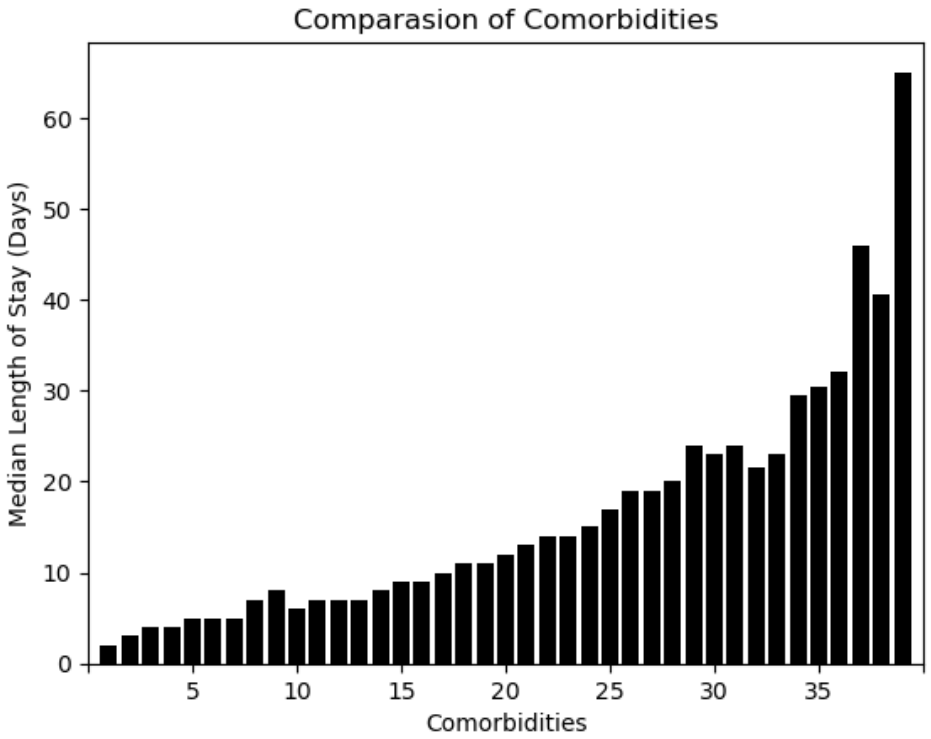

Figure 2. Median LOS depending on the number of comorbidities

Median length of stays depending on the categorical variables (predictors) were calculated by grouping each category subset for comparison (figures 3- 9).

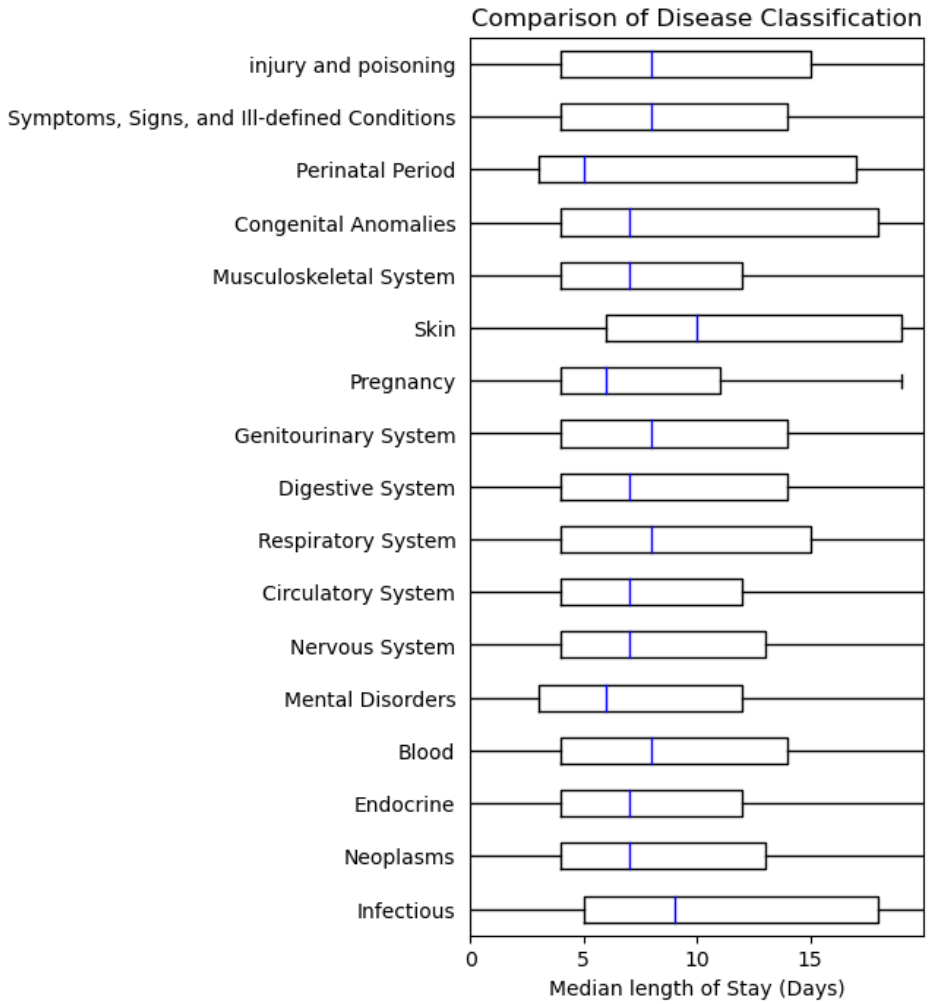

Figure 3. Median LOS depending on the disease classification

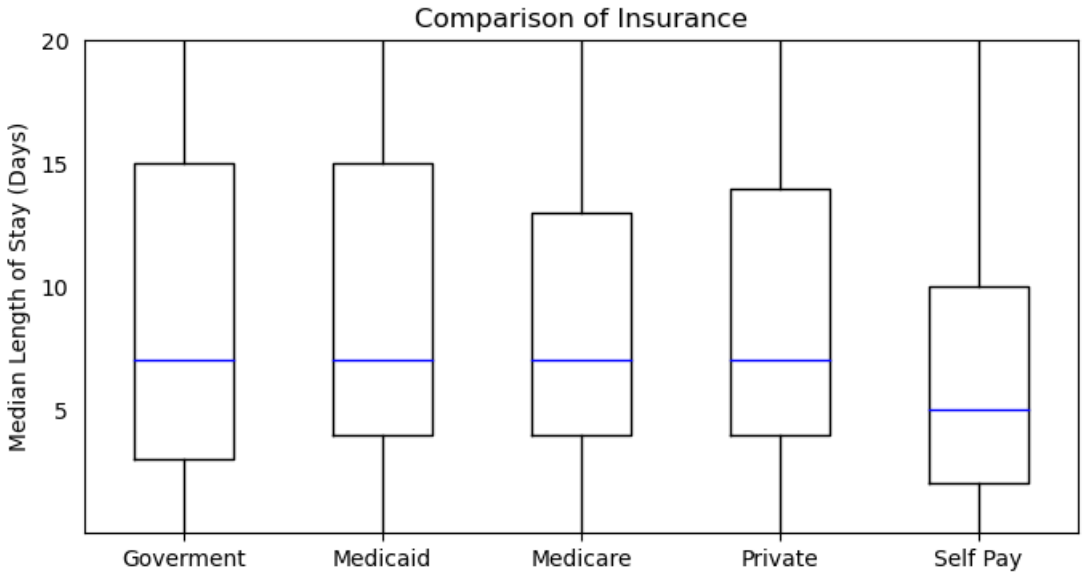

Figure 4. Median LOS depending on the insurance

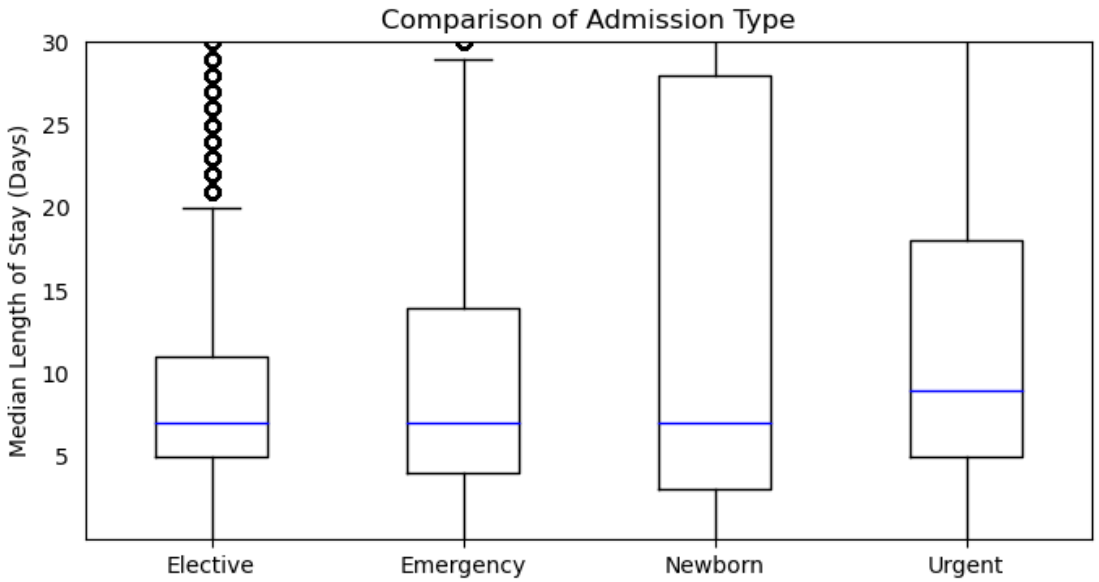

Figure 5. Median LOS depending on the admission type

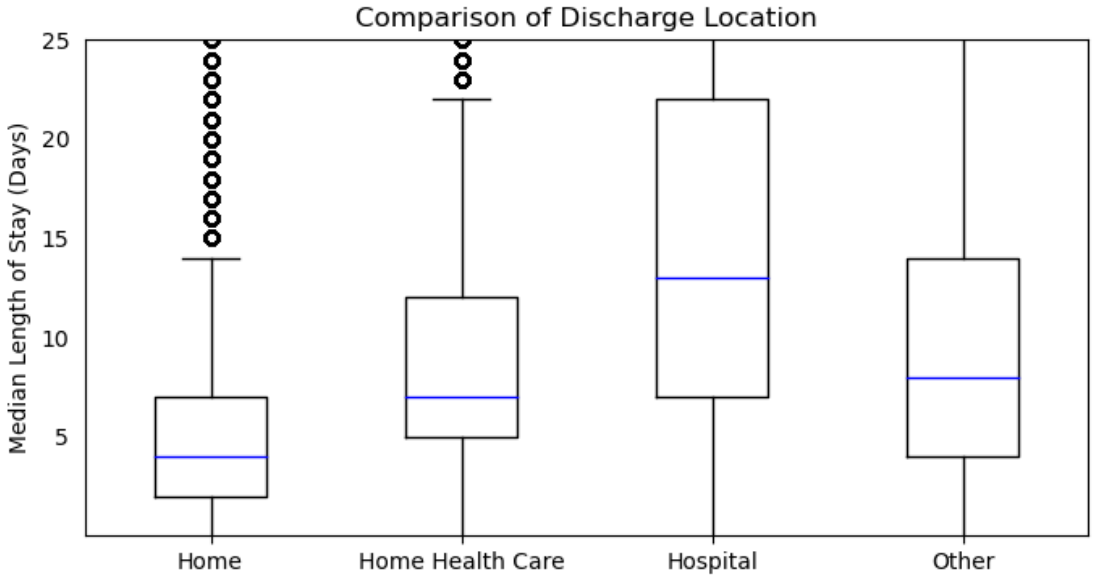

Figure 6. Median LOS depending on the Discharge Location

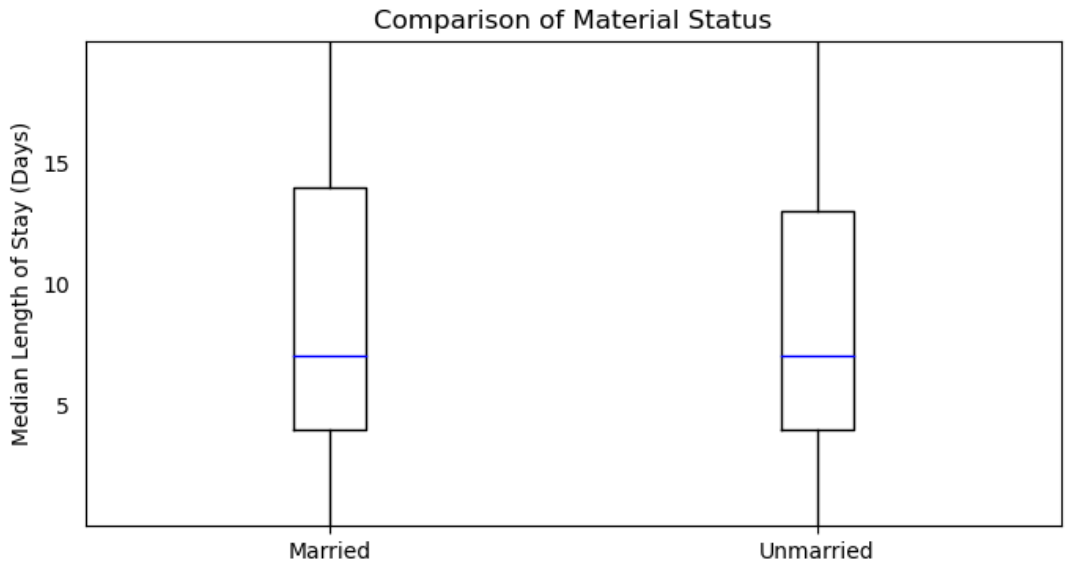

Figure 7. Median LOS depending on the marital status

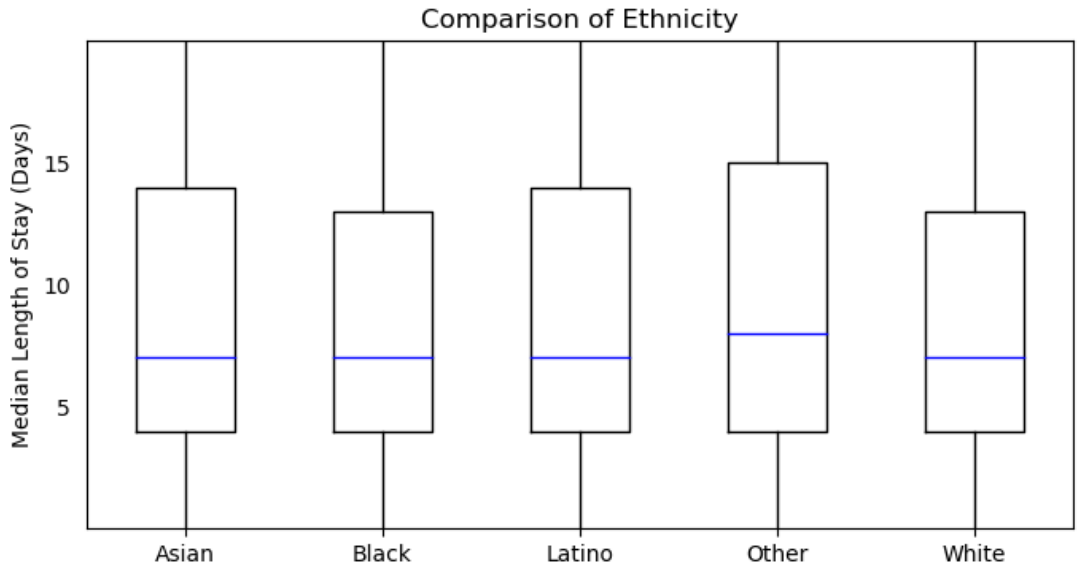

Figure 8. Median LOS depending on the ethnicity

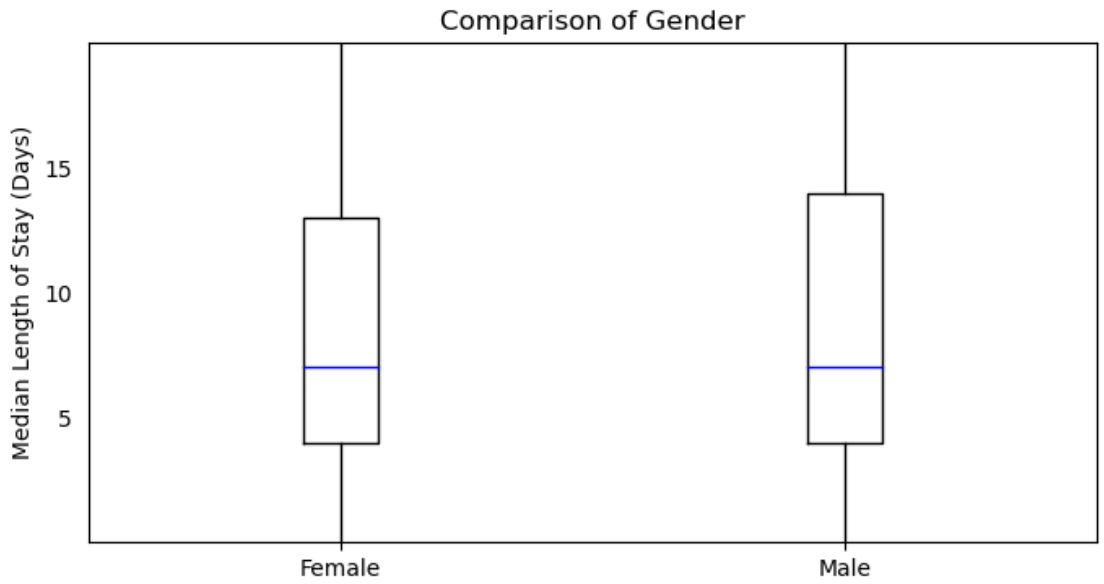

Figure 9. Median LOS depending on the gender

Unnecessary columns and duplicates were removed from the dataset. Then statistical tests were carried out, kruskal-wallis test for categorical data and pearson correlation for numerical value (comorbidities). Alpha was the significance level or threshold chosen to select variables with p-values below this level. Alpha = 0.05 or 5%, p-values less than it, considered as statistically significant. This selection does not guarantee correct decisions with a certain probability because the question to be answered is not about the probability of chance, instead the probability of obtaining an extreme outcome is based on a chance model. Consequently, an evaluation is made about the appropriateness of the chance model, but this judgement does not possess a measurable probability associated with it (Bruce, Bruce and Gedeck, 2020). Hence, features were ranked based on the obtained p-value (Table 1 and 2).

	Features	H	p-value
1	Discharge Location	42335.213836	0.000000e+00
8	Disease Classification	3681.342484	0.000000e+00
3	Religion	836.948120	1.815620e-182
2	Insurance	458.733329	5.620087e-98
0	Admission Type	441.906700	1.848265e-95
7	Age Group	352.370531	4.574867e-76
5	Ethnicity	161.788686	6.043482e-34
6	Gender	76.532524	2.166159e-18
4	Marital Status	72.631115	1.562954e-17

Table 1. Kruska Wallis test results for categorical variables

All the variables (Discharge Location, Disease Classification, Religion, Insurance, Admission Type, Age Group, Ethnicity, Gender, and Marital Status) tested have very low p-values, indicating that there are notable variations in the length of stay of patients in the hospital among the categories of each variables. These results suggest that all the tested variables hold a considerable possibility to act as important predictors in estimating the patients' length of stay.

	Feature	Correlation coefficient	p-value
0	Comorbidities	0.336449	0.0

Table 2. Pearson correlation for numerical value

There is a reasonable positive correlation (r = 0.336) between comorbidities and length of stay in the hospital. Its p-value (0.0) reveals that the correlation is statistically significant, clarifying that the probability of observing such a correlation simply by random chance is doubtful.

3.3 Machine Learning

Random Forest can be defined as a form of ensemble learning that uses trees. It is applied to regression and classification problems (Sekeroglu et al., 2022). The independent variables are evaluated by each tree with binary tests at respective nodes, composing the branches' tree (Moya-Carvajal et al., 2023).

A new dataframe was created with admission type, discharge location, insurance, religion, length of stay, age group, diseases classification and comorbidities. The data was splitted into X which contained the above variables except length of stay (predictors) and Y that included exclusively length of stay (target).

Categorical data present in the X (except comorbidities) were converted into numerical values by applying get_dummies method which created column heading for each subcategory in the categorical variable and encodes as 1 if it matches the category present in the initial categorical column, and assign 0 otherwise.

X and Y were splitted into training (80%) and testing sets (20%) with random_state parameter set to 42. The parameter guarantees that whenever the code is run it uses the same random data split. This splitting technique ensures that machine learning model performance is evaluated on the new data.

Considering that data has values in different ranges, its distribution was standardised by StandardScaler:

$$z = (x - u)/s$$

x is the original value
u is the mean
s is standard deviation
z is the standardised value

Standardisation of features columns in machine learning transforms them to have mean value of 0, and 1 for standard deviation, thus they are more comparable to a standard normal distribution. The process of learning weights is facilitated, enabling more efficient and accurate modelling. Further, standardisation retains crucial information about outliers in the data, making the algorithm less sensitive to them (Raschka, Liu and Dzhulgakov, 2022).

Parameters for the machine learning model were found by the RandomizedSearchCV method that searches through a range of values for specified hyperparameters to identify ideal parameters' combinations. It is a generative process that establishes the parameters to be searched, and samples for evaluation are randomly selected. Random search is the best suited for searching high-dimensional spaces and it outperforms grid search (Mansoori, Zeinalnezhad and Nazarimanesh, 2023). The main parameters to modify are n_estimators, max_features, and pre-pruning configuration such as max_depth. It is generally preferable to choose larger values, using an average of more trees will lead to a more stable and effective ensemble model by minimising overfitting. Nonetheless, there is a point of diminishing returns, as increasing the number of trees will require more memory and training time. There is a possibility that adding max_features or max_leaf_nodes could enhance model performance while simultaneously reducing time and space demands for training and prediction (Müller and Guido, 2016).

The results were applied to create a Random Forest Regression Model. Random Forest is a machine learning algorithm that uses fundamentally an ensemble of decision trees that are slightly different from each other. The concept is that each individual tree could excel on predicting, but may overfit on other parts of the data. If many trees are built, each tree shows good performance and overfitting to different extents, the amount of overfitting can be reduced by averaging their results. Overfitting decrease can be shown through the application of a rigorous mathematical methodology while preserving the predictive capacity of the individual trees. For this approach to be implemented, It is required to build a large number of decision trees that have enough accuracy to predict the target, additionally, each tree must be distinct from the other trees. Random Forest derives from introducing randomness when building trees to assure each tree is unique. Trees in a random forest are randomised by the selection of the data points used to build each tree and the features chosen in each split test (Müller and Guido, 2016).

This model was applied to two different dataset differing on number of features 5 (discharge location, insurance, religion, disease classification and comorbidities) and 7 (age group, admission type, discharge location, insurance, religion, disease classification and comorbidities), respectively (Table 3 e 4). All these features are statistically significant, within the threshold set of less than 0.05 for p-value.

	discharge_location	insurance	religion	length_of_stay	diseases_classification	comorbidities
0	HOSPITAL	Private	UNKONWN	1	Injury and Poisoning	5
2	HOSPITAL	Private	UNKONWN	1	Nervous System	5
3	HOSPITAL	Private	UNKONWN	1	Mental Disorders	5
4	HOSPITAL	Private	UNKONWN	1	Circulatory System	5
5	HOME HEALTH CARE	Medicare	CHRISTIAN	5	Circulatory System	7
...	...	...	...	...	...	...
553791	OTHER	Private	CHRISTIAN	41	Mental Disorders	8
553792	OTHER	Private	CHRISTIAN	41	Endocrine	8
553793	HOME	Private	CHRISTIAN	1	Injury and Poisoning	5
553794	HOME	Private	CHRISTIAN	1	Musculoskeletal System	5
553795	HOME	Private	CHRISTIAN	1	Mental Disorders	5

289093 rows × 6 columns

Table 3. Five features selected to predict the length of stay of patients in the hospital

	admission_type	discharge_location	insurance	religion	length_of_stay	age_group	diseases_classification	comorbidities
0	EMERGENCY	HOSPITAL	Private	UNKONWN	1	Senior	Injury and Poisoning	5
2	EMERGENCY	HOSPITAL	Private	UNKONWN	1	Senior	Nervous System	5
3	EMERGENCY	HOSPITAL	Private	UNKONWN	1	Senior	Mental Disorders	5
4	EMERGENCY	HOSPITAL	Private	UNKONWN	1	Senior	Circulatory System	5
5	ELECTIVE	HOME HEALTH CARE	Medicare	CHRISTIAN	5	Senior	Circulatory System	7
...	...	...	...	...	...	...	...	...
553791	EMERGENCY	OTHER	Private	CHRISTIAN	41	Adults	Mental Disorders	8
553792	EMERGENCY	OTHER	Private	CHRISTIAN	41	Adults	Endocrine	8
553793	EMERGENCY	HOME	Private	CHRISTIAN	1	Senior	Injury and Poisoning	5
553794	EMERGENCY	HOME	Private	CHRISTIAN	1	Senior	Musculoskeletal System	5
553795	EMERGENCY	HOME	Private	CHRISTIAN	1	Senior	Mental Disorders	5

289093 rows × 8 columns

Table 4. Seven features selected to predict the length of stay of patients in the hospital

The performance metrics for model analysis were mean absolute error (mae), mean square error (mse) and coefficient of determination values (r^2).

The data was extracted using PostgreSQL and connected to Python to apply all the above steps.

4. Results

4.1 Model performance evaluation

In this study, two forms of feature selection were carried out, the kruskal-wallis test and Pearson correlation to choose the top 5 and 7 features in a total of 10 features for patients length of stay prediction in the hospital.

Random Forest Regression (RF) model was fit to the final data frame incorporating outliers with 5 and 7 features selected and ran once more after the removal of outliers with a similar number of features. Following these processes, RF model performance evaluation was done by applying testing data, which allowed the comparison between mean absolute error (mae), mean square error (mse) and coefficient of determination (r^2) values (Table 5 and 6).

	Model with 5 features			Model with 7 features		
	mae	mse	r^2	mae	mse	r^2
Random Forest	6.78	136.0	0.25	6.61	124.14	0.31

Table 5. Performance evaluation of the Random Forest model built including outliers in the data frame

The average of the squared differences between the predicted and actual values was measured by mean squared error (mse) to verify how close predictions are to the actual values. The lower the mse values, the closer is the forecast to the true value so Random Forest models with 7 features have better performance than models with 5 features.

Coefficient of determination (r^2) values measure the fitness of the model to the data, ranging from 0 to 1, a perfect fit would be 1. In both RF models, it can be observed that models with 7 have higher r^2 values compared to models with 5 features, which indicates that the RF model with 7 features fits the data with the best performance.

The average absolute difference between the predicted and actual values was measured by mean absolute error (mae). Both models have low values of mae but the Random Forest model with 7 features reached the lowest values which shows that the model makes accurate predictions.

4.2 Feature Importance

A set of top features, admission type, age group, discharge location, disease classification, insurance, religion and comorbidities were used to train the model to predict the length of stay of patients in the hospital. Hence, it was measured the importance of individual features in the model considering its enhancement to the Random Forest model performance for predictions. This approach identified the most significant features in predicting the length of stay of patients in hospital.

In the model trained with 5 features it is seen that the top (4) features are: comorbidities, discharge location, insurance and disease classification (Figure 10). While the model with 7 features has comorbidities, admission type, discharge location and insurance with highest scores (Figure 11). On both models, comorbidities is the most important feature to predict length of stay in hospital.

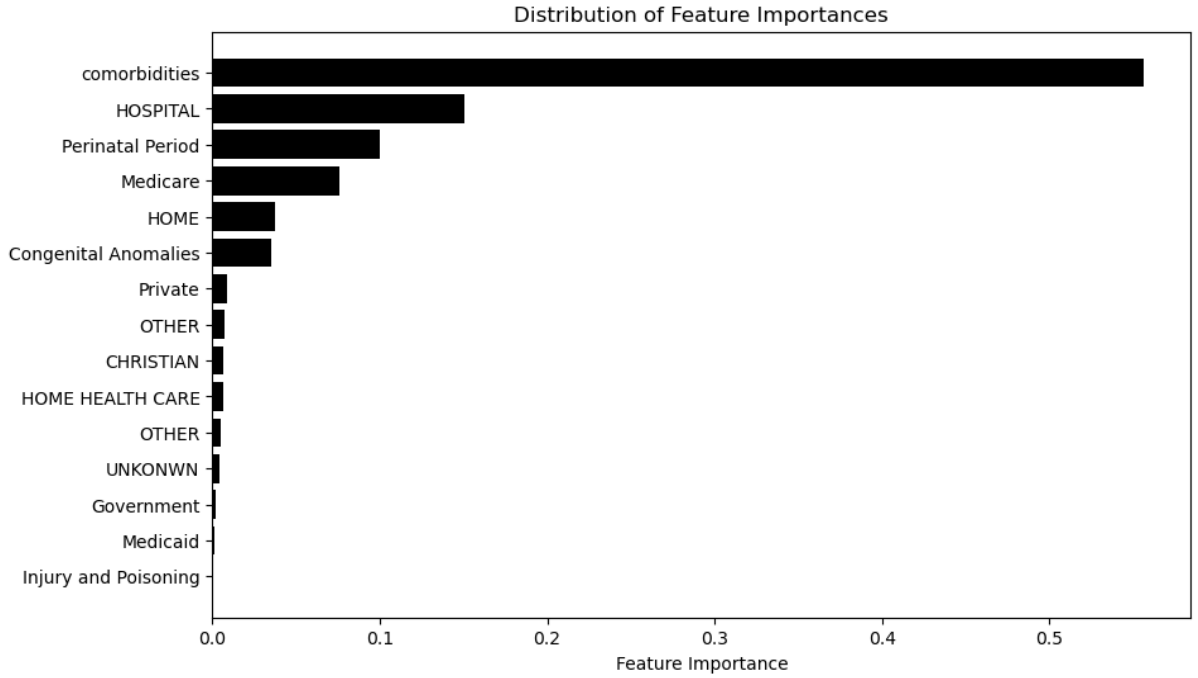

Figure 10. Most important features to predict length of stay in the model with less features (5)

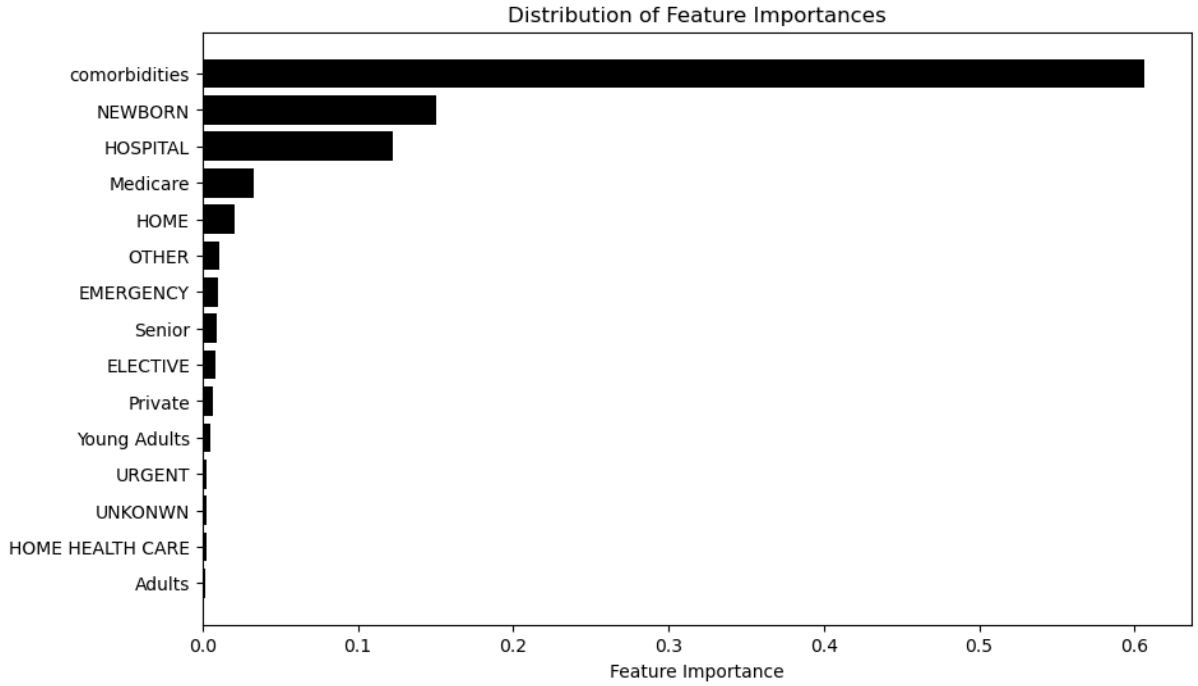

Figure 11. Most important features to predict length of stay in the model with 7 features

Features with low importance scores can be removed from the models and it may affect models' accuracy. In contrast, high importance scores can lead to more collection of data related to the feature so that model performance is improved.

4.3 Predicting length of stay

Visualising predictions results made by the random forest models is an outstanding method to understand models accuracy and efficacy. Scatter plots with predicted and actual length of stay were plotted, patients are represented by the points. Figure 12 and 13 show a positive correlation between the predicted and actual length of stay. Accuracy of the model predictions is defined by the closeness of points to the line so it can be said that the models are performing relatively well.

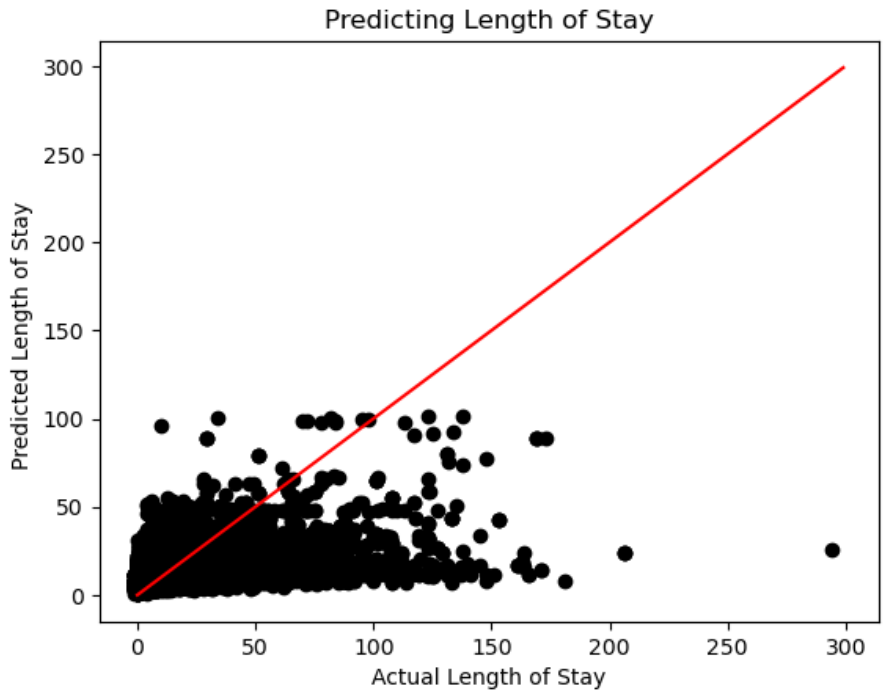

Figure 12. Scatter plots of the actual and predicted length of stay with 5 features

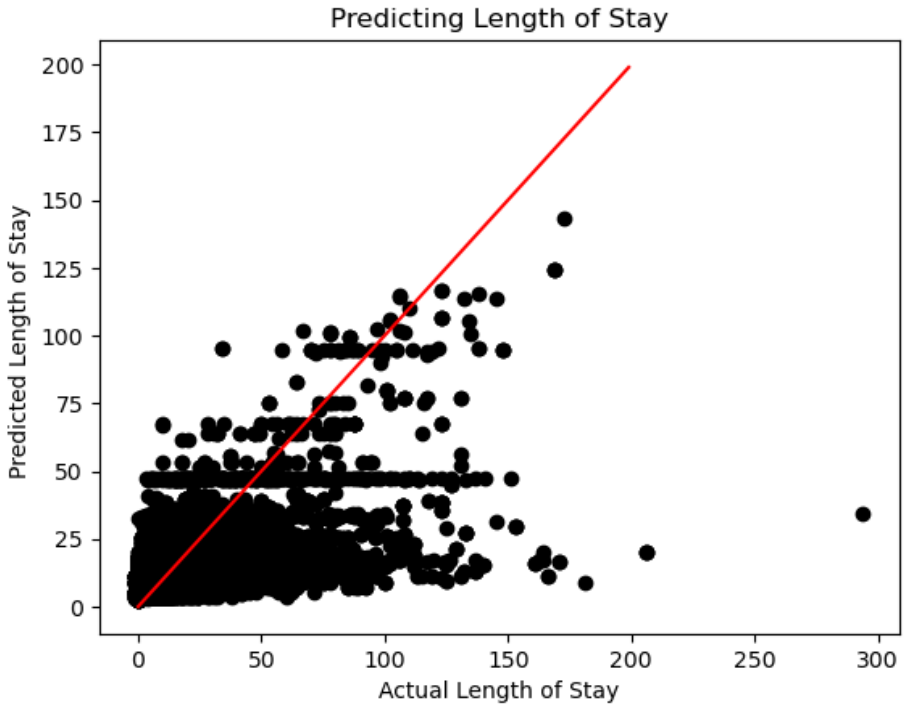

Figure 13. Scatter plots of the actual and predicted length of stay with 7 features

It can be seen that the scatter plots are data points which deviate from the overall pattern, these points can indicate the models' limitation. These outliers can help enhance the model's performance if examined and managed.

Interpretation of graphs making prediction of length of stay (days) permit identity the correlation between predicted and actual length of stay and areas for improvement with data points outliers.

5. Discussion

This study's aim was to create a model able to predict length of stay of patients in the hospital using Random Forest Regression models. Two feature selection approaches were used, Kruskal-Wallis test and Pearson correlation, to select the top 5 and 7 features from 10 potential predictors. The models' performance were evaluated by mean squared (mse), mean absolute (mae), and coefficient of determination (r^2) as well as the assessment of feature importance in the model.

The results from the above analysis show that increasing the number of features in the Random Forest models improves its performance to predict the length of stay of patients in the hospital. The models constituted with seven features performed better than those with five features, it is evidenced by the low values of mse and mae, and high r^2 . In some situations, decisions in the business can depend on the mean squared error or mean absolute error providing model tuning using these metrics. However, in general, coefficient of determination has been identified as a more understandable metric for evaluating regression models (Müller and Guido, 2016). The values of r^2 are interpreted as the fraction of the

information in the data that can be explained by the model. Thus, a coefficient of determination of 0.31 in the model with 7 features in the presence of outliers means that the model can explain 31% of the outcome variation (Table 5). It is an easily interpretable statistic, but r^2 is a measure of correlation, not accuracy (Kuhn and Johnson, 2013).

The most important feature identified in both models was the number of comorbidities in patients which indicates the significance to consider the presence of other diseases for predicting LOS of patients in the hospital. A possible explanation for this conclusion is that patients with various medical conditions could be more vulnerable to developing further complications, as a result, these patients may require extensive medical intervention. Moreover, healthier patients are more independent than those patients with multiple health conditions, leading to prolonged periods of recovery to return to their previous state of wellbeing (Olthof et al., 2014). Other features of importance included discharge location (Home), admission type (Newborn), insurance (Medicare) and disease classification (perinatal period). The significance of individual variables in assessing the model is directly correlated with its relevance in making a prediction, rather than its impact on the model's accuracy (Rezaei-Hachesu et al., 2013).

The visualisation of prediction results using scatter plots exhibited a positive correlation between predicted and actual length of stay (days). However, there were some outliers that deviated from the overall pattern, potentially signalling limitations of the models. Handling these outliers improved the model's performance, as seen when visualising predictions after removal of outliers.

Despite the promising predictive results, there are some limitations in the study. First, this study was conducted within a specific hospital thus the discoveries may not be applicable to a wide range of hospitals and may restrict the models' ability to be implemented in different settings. Furthermore, it was merely used by Random Forest algorithms for predicting length of stay (days) in this study but other machine learning techniques may outperform this algorithm (Orooji et al., 2022).

In general, the examination brings to the forefront importance of the number of features selected and feature importance analysis to forecast the length of stay of patients in the hospital. The results could be used to advance patient treatment and maximise efficient use of hospital resources. Nevertheless, extended analysis is required to validate the models on larger dataset and explore the impact of additional features on model performance.

6. Conclusion

The application of machine learning algorithms using data from the MIMIC III database to predict the length of stay of patients in the hospital show promising results for improving patient care and maximising efficient use of hospital resources. In this study, the Random Forest models were found to be capable in estimating a patient's length of stay (days), the results indicate that increasing the number of features in the model enhanced its

performance. Comorbidities emerged as the most significant predictor of length of stay of patients in the hospital, emphasising the relevance to consider presence of various health conditions in patients. Moreover, the visualisation of the predictions help identify outliers and possible improvement of the model by its removal. Although this study has some limitations, valuable insights are offered by the findings and it can be used to enhance patient treatment and improve allocation of hospital resources. Further investigation with larger datasets and different machine learning techniques is necessary to evaluate the models and examine the influence of additional features on the models perfomance.

7. Acknowledgment

I would like to express my genuine thankfulness to Dr. Marmen Romano for her recommendations and guidance during the progression of the project. Furthermore, it provides an opportunity to consolidate the knowledge regarding Machine Learning techniques and length of stay of patients in Hospital.

8. References

1. Abd-Elrazek, M.A. et al. (2021) “Predicting length of stay in hospitals intensive care unit using general admission features,” Ain Shams Engineering Journal, 12(4), pp. 3691–3702. Available at: https://doi.org/10.1016/j.asej.2021.02.018.

2. Bruce, P., Bruce, A. and Gedeck, P. (2020) Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O’Reilly Media.

3. Delzell, D. a. P. et al. (2019) “Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data,” Frontiers in Oncology, 9. Available at: https://doi.org/10.3389/fonc.2019.01393.

4. Gentimis, T. et al. (2017) “Predicting Hospital Length of Stay Using Neural Networks on MIMIC III Data,” 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech) [Preprint]. Available at: https://doi.org/10.1109/dasc-picom-datacom-cyberscitec.2017.191.

5. Getting Started (no date). Available at: https://mimic.mit.edu/docs/gettingstarted/.

6. Gulati, A.P. (2022) Dealing with outliers using the Z-Score method. Available at: https://www.analyticsvidhya.com/blog/2022/08/dealing-with-outliers-using-the-z-score-method/.

7. Kuhn, M. and Johnson, K. (2013) Applied Predictive Modeling. Springer Science & Business Media.

8. Hu, Z. et al. (2022) "Network analytics and machine learning for predicting length of stay in elderly patients with chronic diseases at point of admission," BMC Medical Informatics and Decision Making, 22(1). Available at: https://doi.org/10.1186/s12911-022-01802-z.

9. Mansoori, A., Zeinalnezhad, M. and Nazarimanesh, L. (2023) "Optimization of Tree-Based Machine Learning Models to Predict the Length of Hospital Stay Using Genetic Algorithm," Journal of Healthcare Engineering, 2023, pp. 1–14. Available at: https://doi.org/10.1155/2023/9673395.

10. Moya-Carvajal, J. et al. (2023) "ML models for severity classification and length-of-stay forecasting in emergency units," Expert Systems With Applications, 223, p. 119864. Available at: https://doi.org/10.1016/j.eswa.2023.119864.

11. Müller, A.C. and Guido, S. (2016) Introduction to Machine Learning with Python: A Guide for Data Scientists. "O'Reilly Media, Inc."

12. Neto, C. et al. (2020) Prediction of Length of Stay for Stroke Patients Using Artificial Neural Networks, Advances in intelligent systems and computing. Springer Nature, pp. 212–221. Available at: https://doi.org/10.1007/978-3-030-45688-7_22.

13. Olthof, M. et al. (2014) "The association between comorbidity and length of hospital stay and costs in total hip arthroplasty patients: a systematic review," Journal of Arthroplasty, 29(5), pp. 1009–1014. Available at: https://doi.org/10.1016/j.arth.2013.10.008.

14. Orooji, A. et al. (2022) "Comparing artificial neural network training algorithms to predict length of stay in hospitalized patients with COVID-19," BMC Infectious Diseases, 22(1). Available at: https://doi.org/10.1186/s12879-022-07921-2.

15. Raschka, S., Liu, Y. and Dzhulgakov, D. (2022) Machine Learning with Pytorch and Scikit-Learn: Develop Machine Learning and Deep Learning Models with Python. Packt Publishing.

16. Sekeroglu, B. et al. (2022) "Comparative Evaluation and Comprehensive Analysis of Machine Learning Models for Regression Problems," Data Intelligence, 4(3), pp. 620–652. Available at: https://doi.org/10.1162/dint_a_00155.

17. Straney, L.D. et al. (2017) “Modelling risk-adjusted variation in length of stay among Australian and New Zealand ICUs,” PLOS ONE. Edited by S. Brakenridge, 12(5), p. e0176570. Available at: https://doi.org/10.1371/journal.pone.0176570.

18. Suha, S.A. and Sanam, T.F. (2022) “A Machine Learning Approach for Predicting Patient’s Length of Hospital Stay with Random Forest Regression,” 2022 IEEE Region 10 Symposium (TENSYMP) [Preprint]. Available at: https://doi.org/10.1109/tensymp54529.2022.9864447.

19. Tsai, P.-F. et al. (2016) “Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network,” Journal of Healthcare Engineering, 2016, pp. 1–11. Available at: https://doi.org/10.1155/2016/7035463.

20. Zolbanin, H.M. et al. (2022) “Data analytics for the sustainable use of resources in hospitals: Predicting the length of stay for patients with chronic diseases,” Information &Amp; Management, 59(5), p. 103282. Available at: https://doi.org/10.1016/j.im.2020.103282.