Dalhousie University

IENG 6964 Optimization of Health Systems

Winter 2022

Predicting the Length Of Stay Using NAÏVE Bayes, KNN and Random Forest Algorithm:

Name: Abilash Surendran
Banner ID: B00891410

Table of Contents:

List of figures:

1. Abstract:

Length of stay is the total number of days a patient has stayed in a hospital. There are several factors that affect the length of stay which include bed grade, the severity of illness, admission type etc. The length of a patient's stay is a key measure of a hospital's management efficiency. Due to limited resources, hospitals must make optimal use of beds and clinician time. The ability to predict how long a patient will stay based on information available as soon as they arrive at the hospital and are diagnosed can have a number of positive consequences for a hospital's efficiency. This research has incorporated three different machine learning algorithms like the Gaussian Naïve Bayes algorithm, KNN algorithm and random forest algorithm. The dataset was taken from Kaggle, a publicly open data source. The train and test data were downloaded as the .csv file and a predictive model was created by incorporating these three different ML algorithms. The model was then trained with the required train dataset. The trained model was then used to predict the LOS in test data. The accuracy of the prediction model is identified and is then compared which each other. The best from the three models are taken.

Keywords: Length of stay, KNN algorithm, naïve gaussian algorithm, Random Forest algorithm

2. Introduction:

Length of stay is a bigger threat Globally. Reduced duration of stay is a key healthcare effort. The entire length of stay from admission to discharge is added for each patient is added to determine the length of stay. The major goal of the management is to lower the patient's total length of stay in order to offer quality treatment. It is one of the variables that may be used to assess the quality of treatment provided to a patient. Longer stays have a detrimental impact on the rate of bed turnover. Predicting the length of stay for patients in emergency, trauma, or urgent situations, on the other hand, is critical because they may have complexities. In addition, a longer stay raises the overall amount of money spent on the patient's treatment.

Trauma care costs more than $37 billion per year in the United States [10]. In Canada, almost 200,000 hospitalizations for trauma are documented each year, costing over 11 billion dollars [7]. Predicting PLOS early may thus assist hospital administration and physicians in mobilising the necessary resources to promote early departure, resulting in improved patient treatment results, cost savings, and increased patient and family satisfaction. Machine learning and deep learning come in handy to solve this issue. They are concepts of statistical learning that can be incorporated in hospital 4.0 to determine and predict the duration of stay.

3. Literature review:

Predicting the length of stay for an inpatient is a difficult and challenging task for a hospital's operational performance. The length of stay is a metric used to assess hospital performance. A lower readmission rate, better resource management, and more efficient services all result from shorter stays. A patient's social difficulties, a lack of services, a lack of facilities, a lack of fault detection equipment, and other issues can all lengthen his or her stay. The LOS is predicted using machine learning methods such as decision trees, Naive Bayes, and k-nearest neighbour algorithms. In this journal, the necessity of choosing the proper qualities was discussed. These models' overall accuracy is measured in percentage deviation.[1]

Clinical hazards and higher expenses are associated with excessively extended hospital stays. Deep venous thrombosis, disuse atrophy, adverse drug reactions, medication blunders, and a variety of additional side events are among the clinical dangers. The severity of complications and duration of stay are directly connected to each other. The length of stay (LOS) in hospitals has been highlighted as a primary driver of resource consumption in a variety of ways. Hospital costs rise as patients take up more beds and staff resources, as well as the number of related adverse events rises.[5]

Predicting the length of stay in the coronary care unit or cardiac intensive care unit for cardiovascular patients is critical. Machine learning algorithms like gradient boosting regression are used to predict. An attempt to predict the LOS is done with patients' historical data from EHR. The traditional regression model performed better than the deep learning model.[2]

4. Methodology:

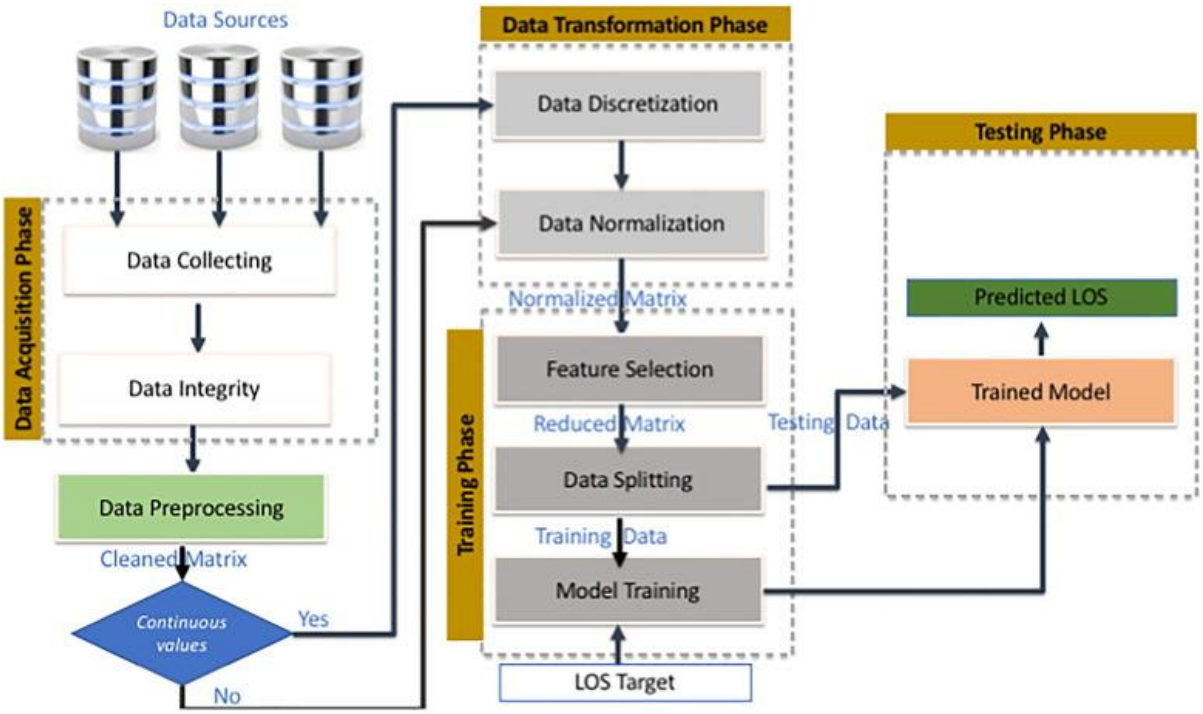

Fig. 1. LOS Prediction Framework. [8]

Machine learning is a technique that incorporates artificial intelligence to predict the desired output. With the revolution of healthcare 4.0 ML algorithms are widely used to classify and predict various factors that affect the quality of care experienced by the patient. The overall length of stay of the patient can be predicted by using various machine learning algorithms. Out of which naïve gaussian, RF and KNN are used in this study.

4.1 Gaussian Naïve Bayes:

Gaussian Naïve Bayes is a type of naïve Bayes that follows a normal distribution and supports continuous data. It just used the mean and standard deviation from the training data to predict the values of the test data. It predicts the data based on a historical train dataset.

4.2 KNN algorithm:

KNN algorithm also called k- nearest neighbour is a simple supervised machine learning algorithm that can be used for the classification of data. Since our data has multiple values, the KNN algorithm would be easy to implement. It relies on labelled input data to learn a function and gives an appropriate output when unknown data is given. The value of k is used to classify the position of new data from the set. If the value of k is 1 it leads to overfitting.

4.3 Random Forest algorithm:

Random forest is a type of ENSEMBLE method. Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. A random forest algorithm is a type of supervised machine learning attribute. It combines the simplicity of a decision tree with flexibility resulting in a vast improvement in inaccuracy. The algorithm creates bootstrapped datasets, that is it randomly selects attributes from the dataset in the training set. Various trees are created using the bootstrapped dataset which provides a greater ensemble to aggregate over. These values to be tested are passed through these trees to predict the optimal solution.

4.4 Dataset:

In order to train the ML algorithm, a vast set of data is required. The data for the current research is obtained from an open source called Kaggle[11]. The test and train data are collected and with respect to obtained data, corresponding algorithms are utilised. The parameters included are case id, hospital code, hospital type code, city code hospital, hospital region code, available extra rooms in the hospital, department, ward type, ward facility code, bed grade, patient id, city code patient, type of admission, the severity of illness, visitors with the patient, age, admission deposit and stay.

4.5 Data Pre-processing:

The required libraries to perform the analytics are identified and are imported in python. The train data sets are mounted to initiate the process. With the help of the panda's function, the test data is read in ML. This function is used to identify the nature of the dataset. The dataset is checked for null values. These are values that decrease the efficiency of the prediction. They are identified and removed at the pre-processing. Data visualisation is done by creating several histograms with the training data. They are listed below.

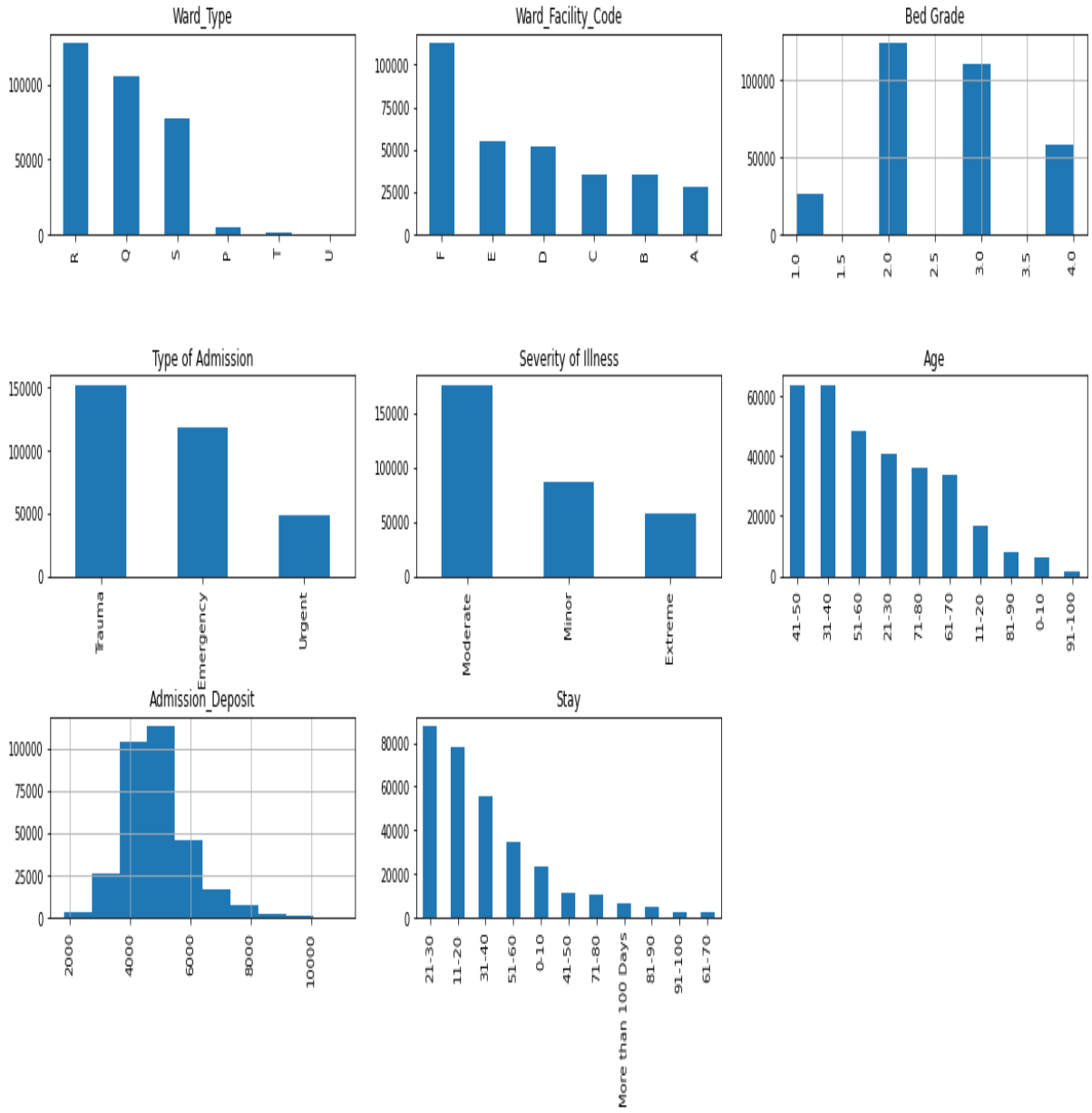

fig 2: data visualisation of input train dataset

The admission type parameter was divided into three categories: emergency, trauma, and urgent, while the severity of illness was divided into three categories: moderate, minor, and extreme. Categorical variables are difficult for the algorithms to read. The categorical data such as the department, hospital area code, ward type, admission type, and severity of illness are encoded using an encoder function. The age and stay (target variable) are spread over a range. They are also encoded to enhance the performance of the algorithm.

4.6 Heatmap:

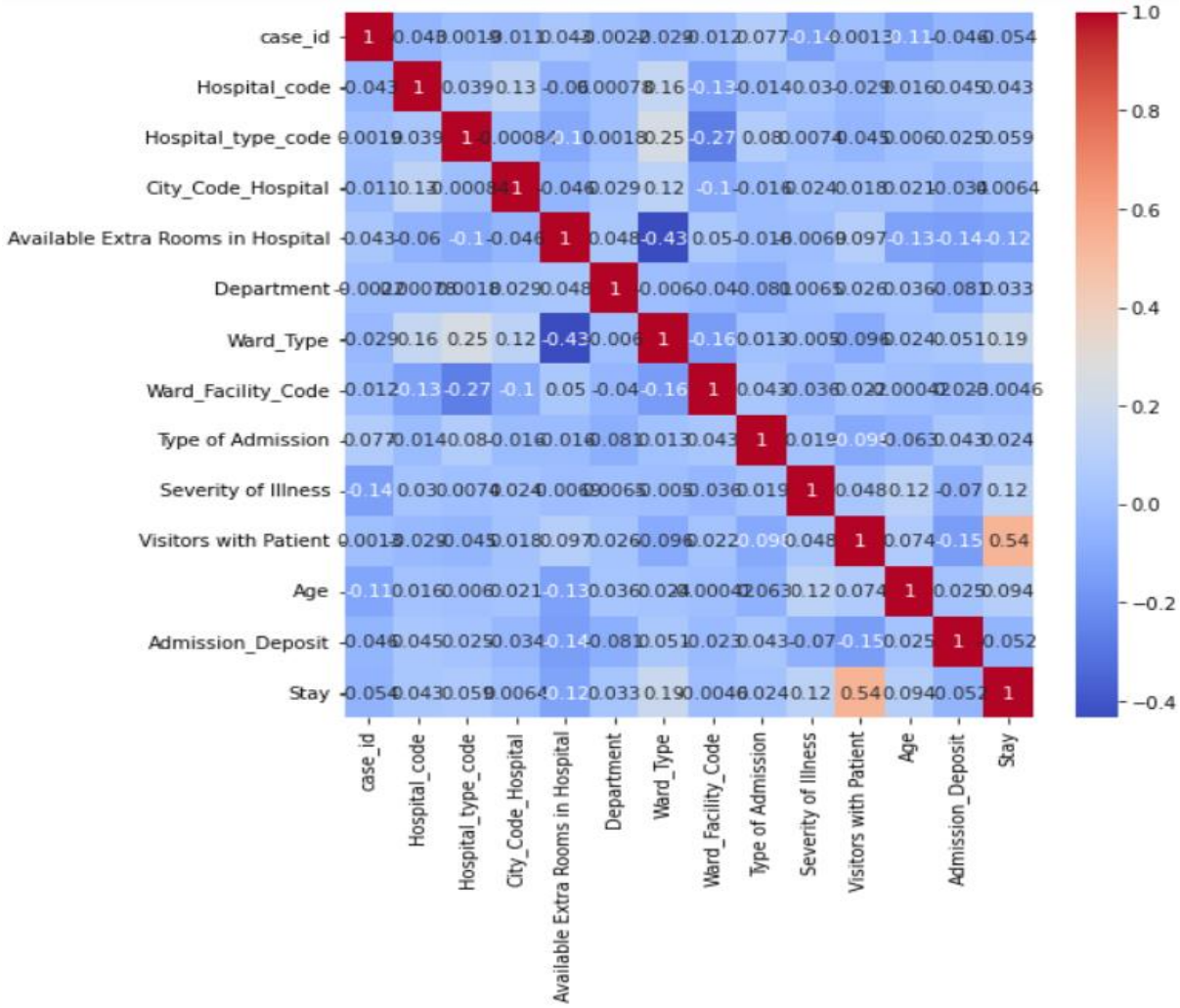

fig 3: heatmap

Examining a spreadsheet full of numbers to find important qualities in a dataset is a time-consuming operation. As a result, Explanatory Data Analysis is used to compile a list of their most important characteristics. The heatmap is a visual representation of the integration of various attributes that are listed in the dataset. It is used to identify the correlation among the different variables within the train and test data. Each square represents the association between the two intersecting variables and aids in the development of predictive models. The map ranges from -0.4 to 1.

4.7 Modelling:

The train data set is combined with the Gaussian naive Bayes, KNN, and random forest algorithms. Before executing the algorithm, non-value-added features found from the heatmap are removed. Before the test data is supplied, the algorithm is trained. The test data is passed through these three algorithms, and the accuracy of their predictions is calculated. For better understanding, these numbers are presented in a bar graph.

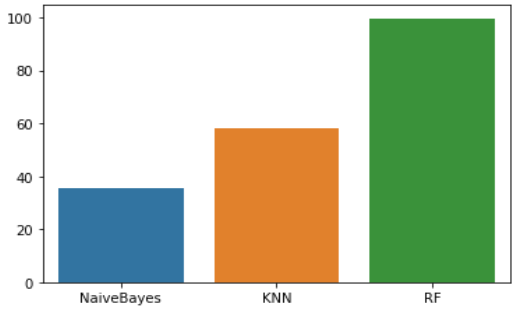

fig 4: Comparison of accuracy

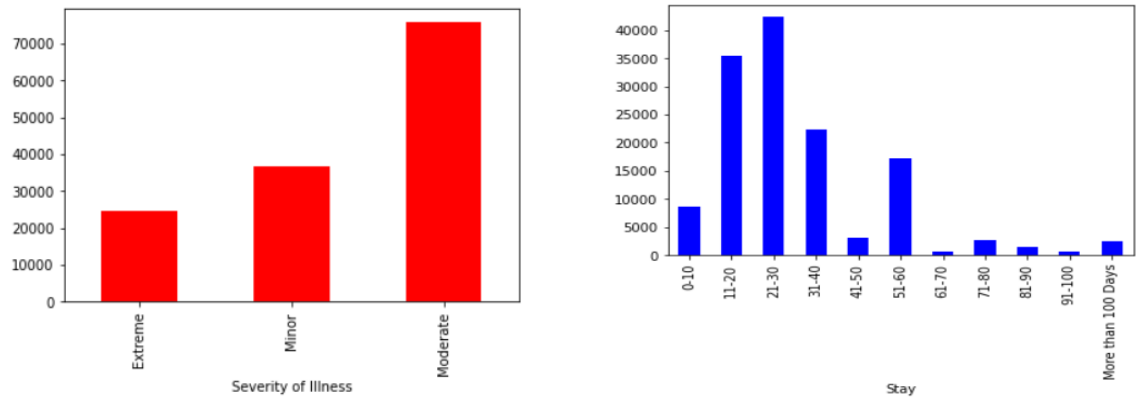

fig 5: The predicted trend for random forest

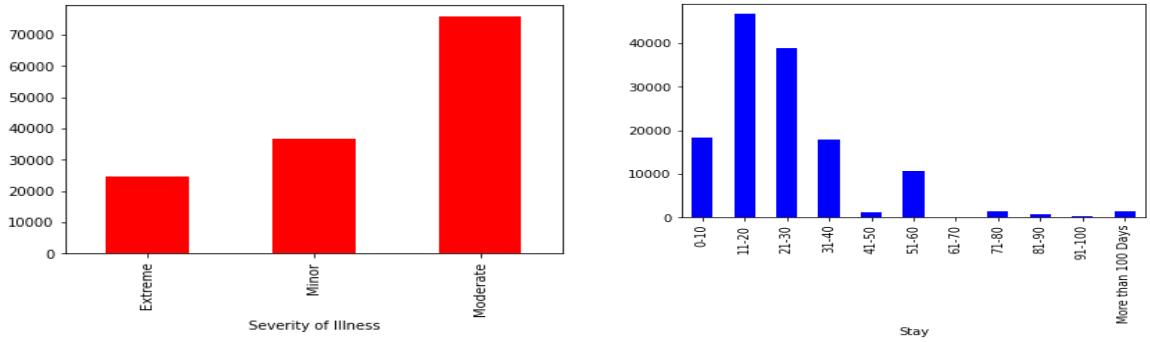

fig 6: The predicted trend for KNN algorithm

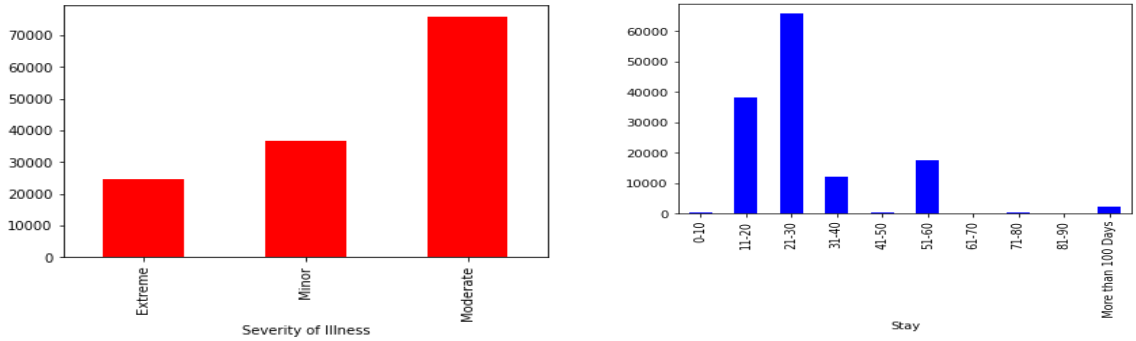

fig 7: The predicted trend for Gaussian Naïve Bayes

The prediction accuracy for the gaussian naïve Bayes algorithm is 35.43%, for the KNN algorithm is 58.11% and for the random forest is 99.71%. The output predicted values are plotted in a bar graph to compare the prediction pattern with the trained data. It is evident that the bar graph of the predicted of LOS by RF algorithm is similar to the that of the train dataset.

5. Conclusion and future scope:

The use of machine learning algorithms in the prediction of length of stay is demonstrated in this study. The RF algorithm outperformed the other two machine learning algorithms. This kind of prediction can be utilised to prevent staying for lengthy periods of time. To deliver high-quality treatment, they can be integrated with EHR and healthcare 4.0. This prediction approach may be used to assign resources for remote patient monitoring in the most efficient way possible. To optimise resource allocation, regression models may be used with these categorization models.

6. Reference:

1. Aghajani, S., Kargari, M. Determining Factors Influencing Length of Stay and Predicting Length of Stay Using Data Mining in the General Surgery Department. *Hospital Practices and Research*, 2016; 1(2): 53-58. doi: 10.20286/hpr-010251
2. Alsinglawi, B., Alnajjar, F., Mubin, O., Novoa, M., Alorjani, M., Karajeh, O., & Darwish, O. (1970, January 1). *Predicting length of stay for cardiovascular hospitalizations in the intensive care unit : Machine learning approach*. Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2020): Enabling Innovative Technologies for Global Healthcare, 20-24 July 2020, Montreal, Canada. Retrieved April 8, 2022, from https://researchdirect.westernsydney.edu.au/islandora/object/uws%3A57369
3. Bahrami N, Soleimani MA, Shraif Nia SH, Shaigan H, Masoodi R, Ranjbar H. Predicted duration of hospital stay and percentage of mortality in patients intensive care unit with APACHE IV. Urmia Med J. 2012;23(5):466-70.
4. Hachesu PR, Ahmadi M, Alizadeh S, Sadoughi F. Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthc Inform Res. 2013;19(2):121-9.
5. Hwabejire JO, Kaafarani HM, Imam AM, Solis CV, Verge J, Sullivan NM, DeMoya MA, Alam HB, Velmahos GC. Excessively long hospital stays after trauma are not related to the severity of illness: Let's aim to the right target. JAMA surgery. 2013 Oct 1;148(10):956-61.
6. Liu P, El-Darzi E, Vasilakis C, Chountas P, Huang W, Lei L. Comparative analysis of data mining algorithms for predicting inpatient length of stay. PACIS 2004 Proceedings. 2004:86.
7. Moore L, Stelfox HT, Turgeon AF, Nathens A, Bourgeois G, Lapointe J, Gagné M, Lavoie A. Hospital length of stay after admission for traumatic injury in Canada: a multicenter cohort study. Annals of surgery. 2014 Jul 1;260(1):179-87.
8. Merhan A. Abd-Elrazek, Ahmed A. Eltahawi, Mohamed H. Abd Elaziz, Mohamed N. Abd-Elwhab, Predicting length of stay in hospitals intensive care unit using general admission features, Ain Shams Engineering Journal, Volume 12, Issue 4, 2021, Pages 3691-3702, ISSN 2090-4479, https://doi.org/10.1016/j.asej.2021.02.018.
9. M.P. Quinn, A.E. Courtney, D.G. Fogarty, D. O'Reilly, C. Cardwell, P.T. McNamee, Influence of prolonged hospitalization on overall bed occupancy: a five-year single-centre study, *QJM: An International Journal of Medicine*, Volume 100, Issue 9, September 2007, Pages 561–566, https://doi.org/10.1093/qjmed/hcm064

10. Velopulos CG, Enwerem NY, Obirieze A, Hui X, Hashmi ZG, Scott VK, Cornwell III EE, Schneider EB, Haider AH. National cost of trauma care by payer status. journal of surgical research. 2013 Sep 1;184(1):444-9.

11. "https://www.kaggle.com/code/abhijeetbhilare/how-long-patient-stay-in-the-hospital/notebook" – source for train and test dataset.