Final Project Report

DAB 402

CAPSTONE PROJECT

LENGTH OF STAY IN EMERGENCY DEPARTMENT

Submitted By:

Student Name	Student ID
Jaspreet Kaur	0730470
Kanchan Bagga	0732356
Varinderjit Singh	0730482

Submitted To:

Pr. Mohammad Shahid (Supervisor)

April 17, 2020

Content

Abstract

This project if all about finding the Length of Stay (LOS) in Emergency Department (ED). This project is made by a group of three students of Data Analytics course in St. Clair College, Windsor, ON, Canada using different machine learning and modeling techniques. This work is done on MIMIC III dataset. This paper describes each task of project related work in very detail from downloading process of data to model building and conclusion. Data analysis is also done to find which diagnosis should be given priority. Five causes for highest death rate are investigated. Many of the machine learning models are trained using Python that predict the LOS. Neural network is one of them. This paper also shows, the neural network gives best accuracy for both training and testing.

The data after cleaning, for training and testing is splitted in the ratio of 80:20 respectively. But for the neural network the data split ratio is 60:40 as it gives best accuracy with this division. All the methods, reasons behind using each if them is described in here.

Length of Stay in Emergency Department

Problem Statement

Today, hospitals are facing the main problem regarding holding the patients to cure mainly in the emergency department because generally patients have an equal waiting time with a guess for the treatment to cure with the physical sickness or any disorder, which is very big problem. To reduce the length of stay in hospital we are using the different techniques for predicting LOS estimation to help the hospitals. Predicting the length of stay improves the hospital planning and manage the resources are used for patients' health benefits and proper care to recover as soon as possible. Through this thing we must be talented to get the approximation of the grave patients for treatment like bed allotment and other things. For example, U.S. hospital stays cost the well-being scheme at the cost $377.5 billion every year and recent Medicare legislation standardizes payments for procedures performed, regardless of the number of days a patient spends in the hospital. We chose this topic because healthcare field was a part of our recent course. Moreover, it will be helpful to improve the healthcare services by reducing the LOS time and improving the affecting features. Hence the waiting time for other patients will also be reduced and thus healthcare system will be improved. Our project can also help in reducing the expenses for patients and hospital management team.

To predict the length of stay we were using the different models in python. These are very helpful to measure the basic concepts of the hospitals related to the general patient's health issues and treatment to recover as soon as it is possible. Because there are various measures and models calculated to ensure the real accuracy of the dataset. The basic network strategies we used to get the accuracy like neural network was the best to achieve the accuracy at good rate of percentage which is 85% of total. Apart from it, the testing and training session of the following dataset also plays appreciate role to get accurate value. The main encounter of these various features and the typical multi-relationships, which are paid to the loss of generic predictive for the long time to keep yourself up to date about the factors related to the field the dataset belongs.

LOS means length of stay which directly describes the time interval of stay for each patient in hospital’s emergency department. This is the time from entry to exit in ED for each patient based on historical data. It depends upon their condition as well like according to the health emergency it will access the period of cure.

Our project is to build a prediction model that tells the LOS with high accuracy for each patient at preadmission time. LOS varies on various factors. Example – Diagnosis, Emergency type (Sudden or by appointment).

> Age, gender, Marital status, Admission type, History of previous admission, Type of treatment are main features. Some other factors also may have impact on LOS that we will be explored in this project.

This project will help in bed allocation and early appointment dates prediction and we can reduce hospital expenses by reducing LOS.

Literature Review

Summary1:

Predicting Hospital Length of Stay Using Neural Networks on MIMIC III Data

In this paper the authors are exploring the study of prediction for length of stay (LOS) for general patients using the MIMIC III database. They have used three different models those are: Support Vector Machine (SVM), Neural Network, and Decision Tree. They have trained a neural network to predict the patient’s stay in hospital in the number of days (how many days a patient will stay). They mentioned that out of three models SVM was the most accurate. Their prediction accuracy is approximately 80% and using linear model. They have mentioned that their database contains more than 50,000 records of people admitted to ICU units for 12 years (2001 to 2012). They have used 28 variables from the dataset.

Reference:

Retrieved 17 April 2020, from https://www.researchgate.net/publication/324177552_Predicting_Hospital_Length_of_stay

Summary 2:

Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network

This article shows the use of the neural network techniques to predict the Length of Stay for patients in a cardiovascular unit with one of three primary diagnoses: heart failure (HF), acute myocardial infarction (AMI), and coronary atherosclerosis (CAS). They have mentioned the variations in length of stay on two factors. One factor is hospital characteristics and other one is patient characteristics. They have explored the data for the National Health Service (NHS) in the United Kingdom. In this article they have mentioned that they have collected total 2,424 admission cases for three diagnoses. 872 heart failure (HF) patients, 572 acute myocardial infarction (AMI) patients, and CAS (coronary atherosclerosis) 933 patients. All these patients are over 65 years. They have analyzed the data from October 1, 2010, and December 31, 2011. Artificial Neural Network (ANN) is used in specific areas, such as cervical cytology and early detection of acute myocardial infarction (AMI). This article shows ANNs are more useful in predicting medical outcomes as compare to logistic regression. Using ANN or linear regression model was able to predict correctly for 88.07% to 89.95% CAS patients at the pre discharge stage and for 88.31% to 91.53% at the preadmission stage. For AMI or HF patients, the accuracy ranged from 64.12% to 66.78% at the pre discharge stage and 63.69% to 67.47% at the preadmission stage.

Reference:

Tsai, P., Chen, P., Chen, Y., Song, H., Lin, H., Lin, F., & Huang, Q. (2016). Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network. *Journal Of Healthcare Engineering*, *2016*, 1-11. doi: 10.1155/2016/7035463

Summary 3:

Predicting hospital admission at emergency department triage using machine learning

This study shows that machine learning can robustly predict hospital admission at emergency department (ED) triage and that the addition of patient history improves predictive performance significantly compared to using triage information alone. In this study it is mentioned that the data was collected for all adult emergency department (ED) visits from March 2014 to July 2017. They have used 972 variables to record each patient visit history. They have used three techniques (Gradient boosting, logistic regression, and deep neural network) for prediction. They have used three dataset types: one for triage information only, second for patient history only, and third for full set of variables. A total of 560,486 patient visits were included in the study. They have used patient's disposition as primary response variable that is encoded in a binary variable (1 = admission, 0 = discharge).

Reference:

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0201016

Summary 4:

Predicting hospital length-of-stay at time of admission

This article shows US hospital stays cost at least $377.5 billion per year on health system. It is mentioned that if we have prior knowledge of LOS, it can help in room and bed allocation planning also. They have used MIMIC dataset to implement the prediction model. they split the LOS target variable and features into training and testing data sets using the ratio of 80 and 20 respectively. Using the training set, they have fit five different regression models and then compared the accuracy on the testing dataset. They have compared the accuracy of Gradient Boosting Regressor model and Random Forest Regressor model. As compare Random Forest Regressor they got more accuracy with Gradient Boosting Regressor on testing dataset. The gradient boosting model RMSE is better by more than 24% (percent difference) versus the constant average or median models. They have used the information such as subject id, hospital admission id, admission date/time, discharge time, and many more. In their dataset they have 58,976 admission events and 46,520 unique patients They have used LOS in days as their target variable.

Reference:

Retrieved 18 April 2020, from https://towardsdatascience.com/predicting-hospital-length-of-stay-at-time-of-admission55dfdfe69598

Summary 5:

Analysis of length of hospital stay using electronic health records: A statistical and data mining approach

In this article they have mentioned that they have used the database of patients admitted to a tertiary general university hospital in South Korea between January and December 2013. They have analyzed the patients according to the three categories. Those categories are descriptive and exploratory analysis, process pattern analysis using process mining techniques, and statistical analysis and prediction of LOS. Right now, EHR information and procedure mining innovation were utilized to break down all occasion logs entered among confirmation and release of the patient. This study helps to find the key factors correlating with duration of hospital stay at the prediction stage. The point of this investigation was to decide a strategy that could be applied to assist medical clinics in dealing with the length of inpatient remain more proficiently. In the data preparation phase, they extracted the EHR log data. Then data cleaning process was performed to extract meaningful analysis results. In the data analysis phase, they have used four types of analysis: LOS performance analysis, LOS analysis of transfer patterns, LOS analysis according to diagnosis, and analysis of long-term hospitalization. At the prediction phase, they identified the main factors correlating with the number of days of stay through data analysis and log-based statistical analysis.

Reference:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5898738/

Summary 6:

Estimating Patient's Length of Stay in the Emergency Department with an Artificial Neural Network

This article also describes the length of stay time interval in the hospital by the general patients to recover their illness effectively. They developed an article that validated the artificial neural network by using the excessive number of patients that was more than the 16000 for the purposes of clinical and operational related to the working of the hospital for take care of the patient. According to their prediction on the length of stay of patients it was the average limit of approximately two hours to spend at the hospital through the training set. The waiting time interval of general and serious patients at the hospitals are very high that creates rush and mess at the hospital that’s why their main motive was to reduce the staying time of these general patients in the hospital by using this dataset along with different models and methodologies, which directly supports their motive and easily defines its idea to the public and patients which was proven very helpful to declines the time rates of patients in the hospital. Moreover, in their research an idea regarding the academic level trauma that tells the care provided to the general patients is very excessive than 42,000 annually.

Reference:

Jesse Wrenn, D. (2020). Estimating Patient’s Length of Stay in the Emergency Department with an Artificial Neural Network. Retrieved 5 March 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560706/

Summary 7:

Forecasting Patient Length Of Stay In An Emergency Department By Artificial Neural Networks

In this study based on the emergency department to reduce the length of stay through the same method attained via a regional university hospital emergency sector in eastern portion of Turkey Which says that the general health of patients over there for the treatment to recover their illness by measuring it through testing and training sections of various models to get the accurate value about the accuracy to assist the patients in the hospital and how they are treated and cured at the hospital to fight with their illness to get well as soon as possible by providing the treatment via hospital staff like doctors, nurses, medical equipment’s and their belongings. This data is used to shorten the time of the service provided at hospital in order to decline the time stay of patients to get well from the problems and stay healthy for further times of life, in this process the public plays an equal role as the doctors to cooperate properly for getting good service for their recovery. It also relates to conditions of the patients that how much the illness they have by which they can get the proper medical treatment for the health benefits. that obliges roughly a typical of 40.000 patients per annum base. For example, they gather a whole information of 1500 ED patients who were preserved in the sector in October and November 2010.

Reference:

(2020). Retrieved 5 March 2020, from https://www.researchgate.net/publication/283163617_Forecasting_patient_length_of_stay_in_an_emergency_department_by_artificial_neural_networks

Summary 8:

Recursive neural networks in hospital bed occupancy forecasting

This study depicts the information about the prediction of hospital bed allocation through the recursive neural networks. However, productive arranging of emergency clinic bed use is the essential condition to limit the medical clinic costs. In the gave work we deal with the issue of occupancy estimating in the size of while, with the focus on the personal vacations arranging. They develop a model prediction through the recursive neural networks, which plays out an inhabitance forecast applying supportable confirmation and release information united with outside variables, For example, open and school occasions. The model requires no close to home data on patients or staff. It was streamlined for 60 days (May-September) conjecture throughout the late spring.

The average error was 6.24% computed by the basis of 8 validation sets through an absolute percentage error (MAPE). The projected machine learning model has shown to be viable to standard time-series predicting models and can be suggested for integration in medium-size hospitals automatized arrangement and conclusion making.

Reference:

Kutafina, E., Bechtold, I., Kabino, K., & Jonas, S. M. (2019). Recursive neural networks in hospital bed occupancy forecasting. *BMC medical informatics and decision making*, *19*(1), 39.

Summary 9:

Neural Network Prediction of ICU Length of Stay Following Cardiac Surgery Based on Pre-Incision Variables

This Neural Network prediction is based on the Cardiac Surgery belongs to the Pre-Incision variables which describes the rate of surgical patients amount in the hospital for recovery and the time span they stayed at hospital. According to the provided dataset it was clear that the number of patients gained through this article was excessive who admitted in the hospital by following the various training and testing sessions of the models related to the dataset. Thirty-six variables collected from 185 cardiac surgical patients were analyzed for contribution to ICU length of stay. The Automatic Linear Modeling (ALM) module of IBM-SPSS software recognized 8 influences with statistically substantial relatives with ICU LOS these influences were also investigated with the Artificial Neural Network (ANN) module of the same software. The biased contributions of each factor were then functional to facts for a “new” persistent to foresee ICU length of stay for that separate. Artificial neural networks established a 2-fold better precision than ALM in estimate of experimental ICU length of stay. This superior precision would be supposed to consequence from the volume of artificial neural networks to detention nonlinear properties and advanced order relations. Analytical exhibiting may be of value in initial expectation of dangers of post-operative disease and application of emergency department conveniences.

Reference:

LaFaro, R. J., Pothula, S., Kubal, K. P., Inchiosa, M. E., Pothula, V. M., Yuan, S. C., ... & Perline, R. (2015). Neural network prediction of ICU length of stay following cardiac surgery based on pre-incision variables. *PLoS One*, *10*(12).

Summary 10:

A Neural Network Analysis of Treatment Quality and Efficiency of Hospitals

In most hospitals the basic issue is regard to the quality and the efficiency of the treatment in the hospitals for the general health of the patients, who admitted over there for the treatment to recover the health issues for which we use this model named Neural Network to find the accurate accuracy value of the testing and training session about this dataset. And this model proven very helpful to get appropriate accuracy value which was eighty five percent and this value is most proficient for this dataset shows the healthcare data for the years 2009-2012 were downloaded from the Statewide Planning and Research Cooperative System (SPARCS) of the New York State Department of Health (NYSDOH). According to that articles they conclusions show that there are substantial alterations in length of stay and death rates liable on the handling technique. Dealing outcome shows a robust suggestion with technique and with the patients’ nature upon release. Remarkably, under comparable health environments, patients who are under the public healthcare system tend to have longer length of hospital stays than others. At the end they help us to offer a selection of features to be measured in estimating persistent health outcomes from hospitalization. They show the most important thing like to utilization of the treatment things to use in the hospitals reliable for the serious patients because if they face the more problems, it becomes the very bad for the society.

Reference:

(2020). Retrieved 5 March 2020, from https://www.hilarispublisher.com/open-access/a-neural-network-analysis-of-treatment-quality-and-efficiency-of-hospitals-2157-7420-1000209.pdf

Ethical Concerns

Consent:

As this dataset is opensource, we don’t need to have any consent to collect this dataset.

Consistency:

We have huge volume of records in our dataset. It is very reasonable for experimenting on this dataset. Moreover, the accuracy of different models will be comparable and consistent.

Clarity:

This data is used for model building with highest accuracy to predict length of stay (LOS) for general patients in hospital. So, it is very clear that how we use this data.

Control:

Although, this dataset is publicly available, but it is controlled by the GitHub. Now we have access to this data, but we can use it only for the experimenting purpose. We cannot manipulate it for public but can make some changes (cleaning) according to our requirements only.

Consequences:

This data collection can never harm any individual. Instead, it will help in the better caring in hospital by predicting length of stay. The experiments on this dataset are going to help a lot to hospital admission management team.

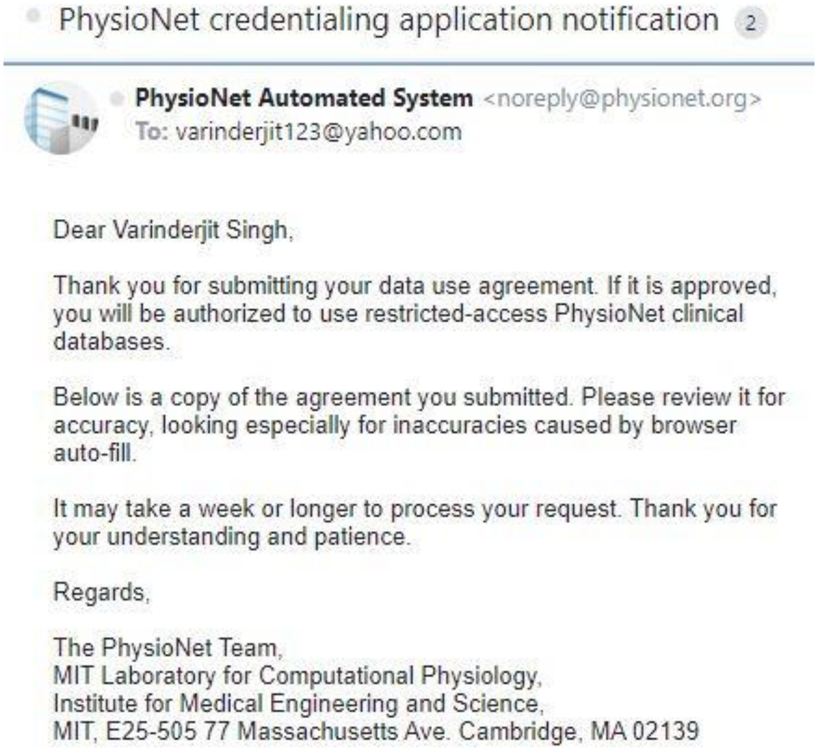
Dataset

Our data is MIMIC-III (**M**edical **I**nformation **M**art for **I**ntensive **C**are III). It is a large, freely available database comprising deidentified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between June 2001 and October 2012.

This database includes information such as demographics, vital sign measurements made at the bedside, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital). The MIMIC-III Clinical Database is available on PhysioNet. We can get more information from their official website.

The process of getting access to database:

First of all, when we were looking for other projects related to LOS, we noticed that most of those projects were made using MIMIC iii dataset. The reason was that this dataset contains very detailed, accurate, well contained, large number of patients, and diseases data. Secondly, we requested to PhysioNet for this dataset by an registered e-mail. When it was received, they replied:



PhysioNet credentialing application notification 2

PhysioNet Automated System <noreply@physionet.org>
To: varinderjit123@yahoo.com

Dear Varinderjit Singh,

Thank you for submitting your data use agreement. If it is approved, you will be authorized to use restricted-access PhysioNet clinical databases.

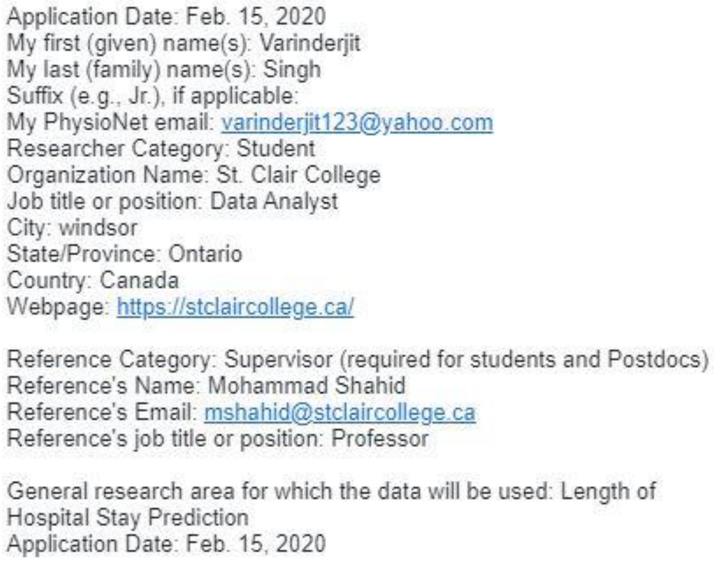
Below is a copy of the agreement you submitted. Please review it for accuracy, looking especially for inaccuracies caused by browser auto-fill.

It may take a week or longer to process your request. Thank you for your understanding and patience.

Regards,

The PhysioNet Team,
MIT Laboratory for Computational Physiology,
Institute for Medical Engineering and Science,
MIT, E25-505 77 Massachusetts Ave. Cambridge, MA 02139

Reference of supervisor was provided:

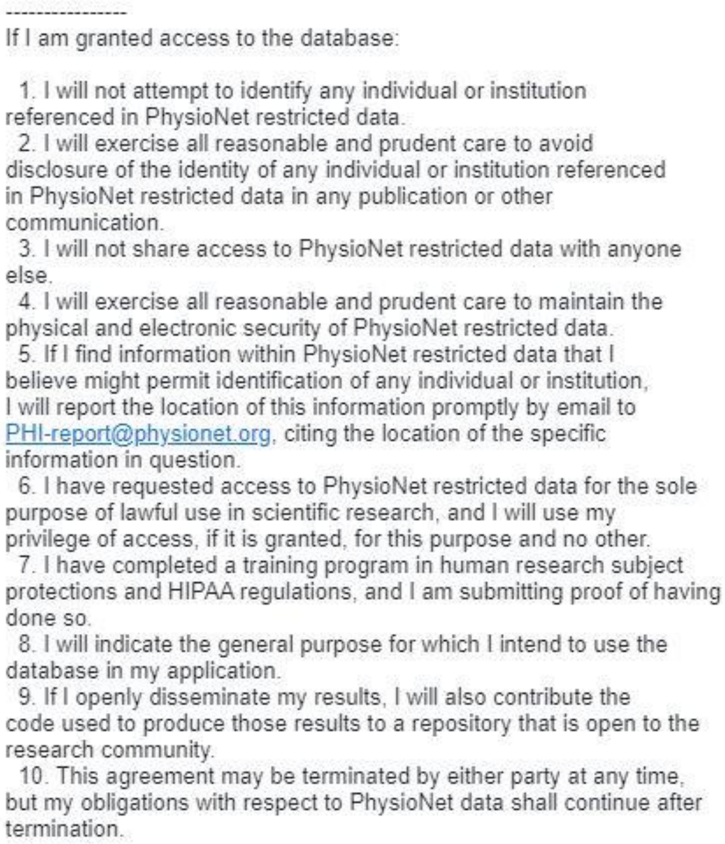


Application Date: Feb. 15, 2020
My first (given) name(s): Varinderjit
My last (family) name(s): Singh
Suffix (e.g., Jr.), if applicable:
My PhysioNet email: varinderjit123@yahoo.com
Researcher Category: Student
Organization Name: St. Clair College
Job title or position: Data Analyst
City: windsor
State/Province: Ontario
Country: Canada
Webpage: https://stclaircollege.ca/

Reference Category: Supervisor (required for students and Postdocs)
Reference's Name: Mohammad Shahid
Reference's Email: mshahid@stclaircollege.ca
Reference's job title or position: Professor

General research area for which the data will be used: Length of Hospital Stay Prediction
Application Date: Feb. 15, 2020

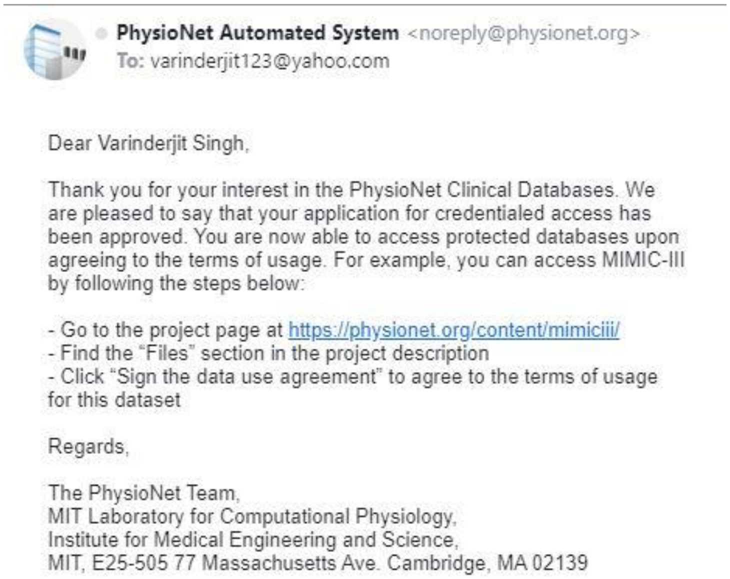
Some promises before getting access:



If I am granted access to the database:

1. I will not attempt to identify any individual or institution referenced in PhysioNet restricted data.
2. I will exercise all reasonable and prudent care to avoid disclosure of the identity of any individual or institution referenced in PhysioNet restricted data in any publication or other communication.
3. I will not share access to PhysioNet restricted data with anyone else.
4. I will exercise all reasonable and prudent care to maintain the physical and electronic security of PhysioNet restricted data.
5. If I find information within PhysioNet restricted data that I believe might permit identification of any individual or institution, I will report the location of this information promptly by email to PHI-report@physionet.org, citing the location of the specific information in question.
6. I have requested access to PhysioNet restricted data for the sole purpose of lawful use in scientific research, and I will use my privilege of access, if it is granted, for this purpose and no other.
7. I have completed a training program in human research subject protections and HIPAA regulations, and I am submitting proof of having done so.
8. I will indicate the general purpose for which I intend to use the database in my application.
9. If I openly disseminate my results, I will also contribute the code used to produce those results to a repository that is open to the research community.
10. This agreement may be terminated by either party at any time, but my obligations with respect to PhysioNet data shall continue after termination.

Finally, we got access:



PhysioNet Automated System <noreply@physionet.org>
To: varinderjit123@yahoo.com

Dear Varinderjit Singh,

Thank you for your interest in the PhysioNet Clinical Databases. We are pleased to say that your application for credentialed access has been approved. You are now able to access protected databases upon agreeing to the terms of usage. For example, you can access MIMIC-III by following the steps below:

- Go to the project page at https://physionet.org/content/mimiciii/
- Find the “Files” section in the project description
- Click “Sign the data use agreement” to agree to the terms of usage for this dataset

Regards,

The PhysioNet Team,
MIT Laboratory for Computational Physiology,
Institute for Medical Engineering and Science,
MIT, E25-505 77 Massachusetts Ave. Cambridge, MA 02139

MIMIC-III is a relational database consisting of 26 tables. The data files are distributed in comma separated value (CSV) format. This dataset contains all the records of hospital.

Now, as our project is for Emergency Department (ED) only, we need the data only for ED. We do not need unnecessary data because its not related to our project. So, we picked data related to just ED and made a single new CSV file that contains only needed columns.

Data Cleaning:

For data cleaning we used Excel because it provides very good data visualization. For the better understanding with data, visualization if data is very important fact.

As, our original dataset had very large data, we picked only the necessary columns. We also needed to create some columns from our side for example:

LOS in ED = ED_AdmTime – ED_DischTime

Diagnosis Code are given to each unique diagnosis because our problem was regression problem so, we needed numeric values only, but original data ICD codes were containing Alpha numeric values.

Null Values are removed.

So now, for this project we have 7543 observations for unique patients. We have 17 columns that contain none of the missing or NULL value. It contains data for 598 unique diagnosis.

These 17 columns are: Serial Number, Subject ID, Hospital Admission ID, Hospital Admission Time, Hospital Discharge Time, Length Of Stay In Hospital, Admission Type name, Admission Type code, Emergency Department Admission Time, Emergency Department discharge Time, ED Stay In Seconds, ED Stay In Minutes, Length Of Stay in ED, Diagnosis, Diagnosis Code, Patient Death Flag, and ICD9 Codes.

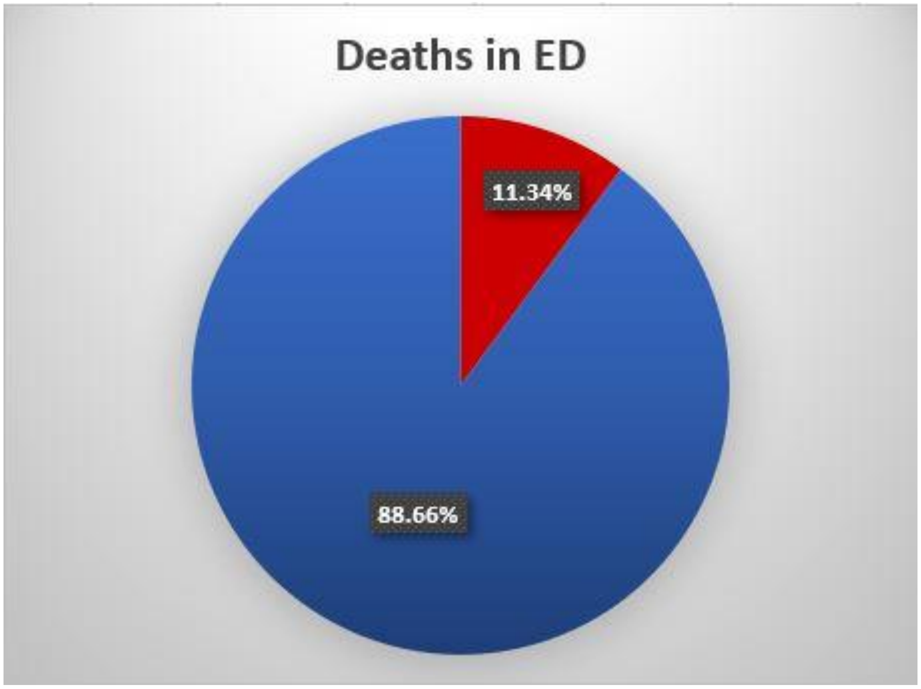
As our dataset was too large for our machines, we chose data only for top 598 unique diagnosis sorted by A-Z alphabetically.

We have 3 types of emergency admissions:

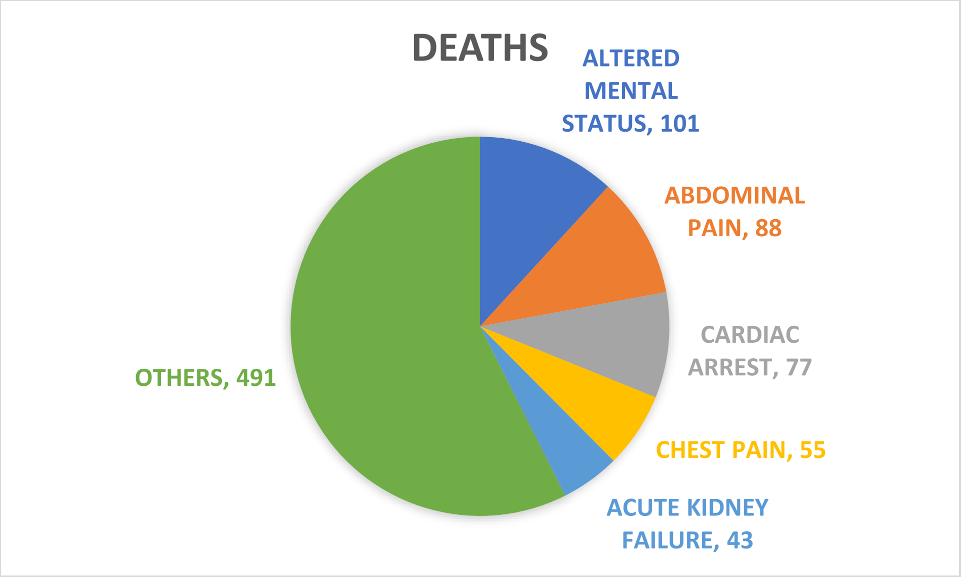
1. Emergency: - Sudden new entry with seriousness.
2. Urgent: - Patient is already admitted in hospital but need some emergency treatment.
3. Elective: - Called by an appointment.

Data Analysis:

Our data shows 855 out of total 7543 patients died in ED. That is 11.34% death rate in ED.



When We analyzed these deaths in deep, we got:



The above pie chart shows the 5 diseases with max death rate in hospital. It clearly shows that max deaths are seen in ED with:

1) Altered Mental Status that is 101 out of 855
2) Abdominal Pain that is 88 out of 855
3) Cardiac Arrest that is 77 out of 855
4) Chest Pain that is 55 out of 855
5) Acute Kidney Failure that is 43 out of 855
6) All other 593 diseases caused 491 deaths from total.

Methods

Modeling techniques we have used:

In this project we have used the different models to get the best accuracy to make the prediction because at the end we need to find the best model to get the best accuracy to predict the length of stay in hospitals for general patients in emergency department.

K-Nearest Neighbors

This method is the very simplest method in machine learning techniques. It is used for both learning like supervised and unsupervised method. In our dataset we predict the length of stay so our target variable is the length of stay in days so that’s why we used that technique as the continuous variables. In the continuous variable it is works as the average like we work on the different numbers of variables then predict as the majority level such as average. In this method used the object is classified by a majority level of its neighbors. For regression purpose to predict the continuous value that values is the average of the values of its l nearest neighbors. For Our dataset MIMIC-III when we use that methods the accuracy of training (99.44%) and testing (99.44%) set is the overfitting model so that model is not efficient for our project to prediction.

Random Forest

Random forest is another method used in our project. It also works on the classification and regression techniques. It utilizes bagging and features randomness when assembling every individual tree to attempt to make an uncorrelated forest of trees whose forecast by board is more precise than that of any individual tree. We used this method because this algorithm contains numerous decision trees. A random forest is a meta estimator that fits various grouping numbers of trees on different subtests of our dataset and utilizations averaging to improve the predictive the length of stay in hospitals and controls the overfitting. But when we applied it on our dataset it shows overfitting on splitting the data 80-20 when we perform again splitting like 70-30 and 60-40 then same result could be produce like on the training (95.43%) and test (92.88%) set the accuracy not good as our expectation.

Gradient Boosting

Gradient boosting is another technique in machine learning for supervised and unsupervised like regression and classification problems. It is used for the prediction in the form of an ensemble of weak prediction models like decision tree. When we read the article on this topic we find the information regarding that model does work very well because it can be performs on the dataset on the basis of it was classifier and regressor which minimal efforts has been spend on the cleaning and learn the more difficult such as complex non-linear decision boundaries through boosting. It generalizes models by optimizing of an arbitrary differentiable function. The accuracy after using the gradient boosting was on training set 95.27% and on test set 95.22%. According to this accuracy that model also not good it depicts the overfit our model.

Decision Tree

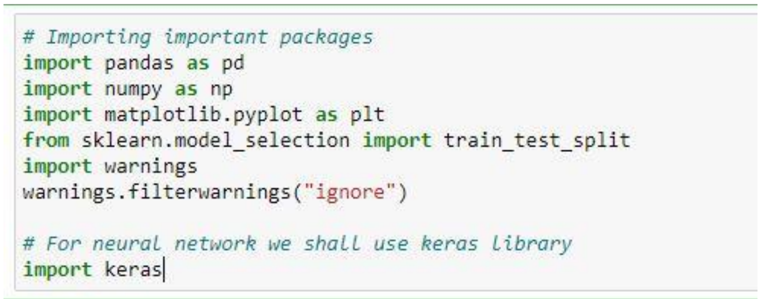
Decision tree is one of the predictive models used in machine learning and data mining. Decision tree are a non-parametric supervised learning method use for classification and regression. Whenever, our target variable is the continuous, so regression technique used by decision tree called regression trees. CART is the general term for regression and classification. To using this method was easy to use and understand for us. It is resistant to outliers just require the little data preprocessing. A decision tree is a decision support tool that uses a tree-like model of decision and their possible consequences, it works according to logical conditions. It is one way to display an algorithm that only contains conditional control statement. The accuracy from through model training set is 80.96% and test set is 80.96%. To see this accuracy, it is also not better for make prediction.

Neural Networks

Neural network is a series of algorithms get together of straightforward preparing components, units or hubs, whose usefulness is inexactly founded on the creature neuron. The handling capacity of the system is put away in the inter unit association qualities or loads acquired by a procedure of adjustment to or gaining from a lot of preparing designs. There are many types of the neural networks in machine learning. In our project we used the MLP neural network (multi-layer perceptions). In our last course we done the machine learning as the subject so we know about the MLP for this project to using the neural networks we read the many tutorials and articles but we did not get the proper information so we used that type to predict the length of stay. As per our information from our book and get the more knowledge about the criteria. At the end This was the best model in our project to get best accuracy on training and testing set.

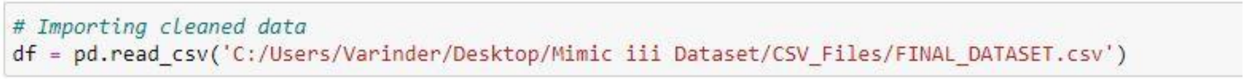
Working with Python

Importing necessary packages:



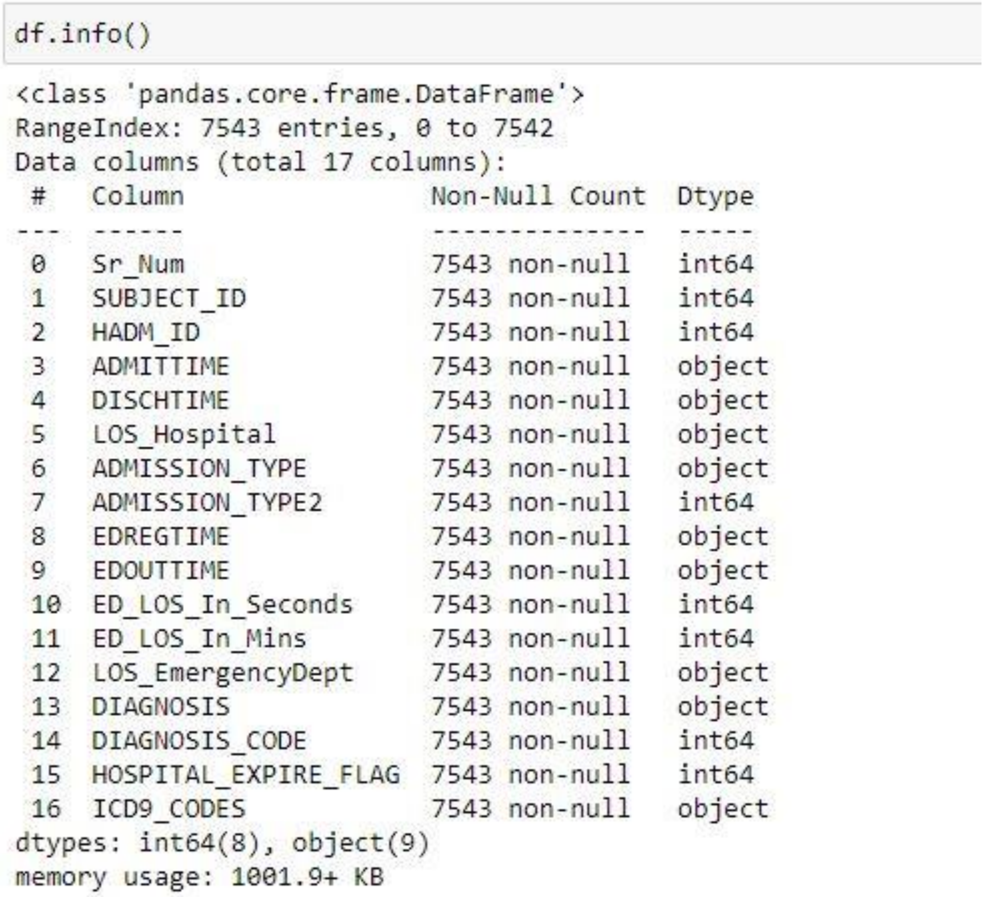
```
# Importing important packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import warnings
warnings.filterwarnings("ignore")

# For neural network we shall use keras library
import keras
```

Importing dataset:



```
# Importing cleaned data
df = pd.read_csv('C:/Users/Varinder/Desktop/Mimic iii Dataset/CSV_Files/FINAL_DATASET.csv')
```

Information about data:

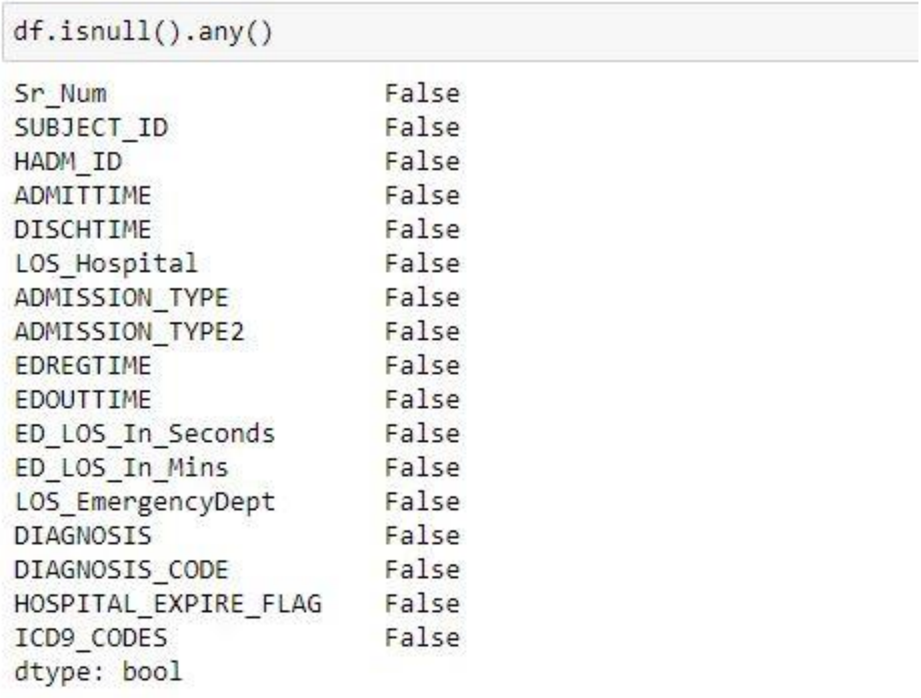


```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7543 entries, 0 to 7542
Data columns (total 17 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   Sr_Num               7543 non-null   int64 
 1   SUBJECT_ID           7543 non-null   int64 
 2   HADM_ID              7543 non-null   int64 
 3   ADMITTIME            7543 non-null   object
 4   DISCHTIME            7543 non-null   object
 5   LOS_Hospital         7543 non-null   object
 6   ADMISSION_TYPE       7543 non-null   object
 7   ADMISSION_TYPE2      7543 non-null   int64 
 8   EDREGTIME            7543 non-null   object
 9   EDOUTTIME            7543 non-null   object
 10  ED_LOS_In_Seconds    7543 non-null   int64 
 11  ED_LOS_In_Mins       7543 non-null   int64 
 12  LOS_EmergencyDept    7543 non-null   object
 13  DIAGNOSIS            7543 non-null   object
 14  DIAGNOSIS_CODE       7543 non-null   int64 
 15  HOSPITAL_EXPIRE_FLAG 7543 non-null   int64 
 16  ICD9_CODES           7543 non-null   object
dtypes: int64(8), object(9)
memory usage: 1001.9+ KB
```

We have 7543 observations in our dataset

Confirming if NULL values are there:

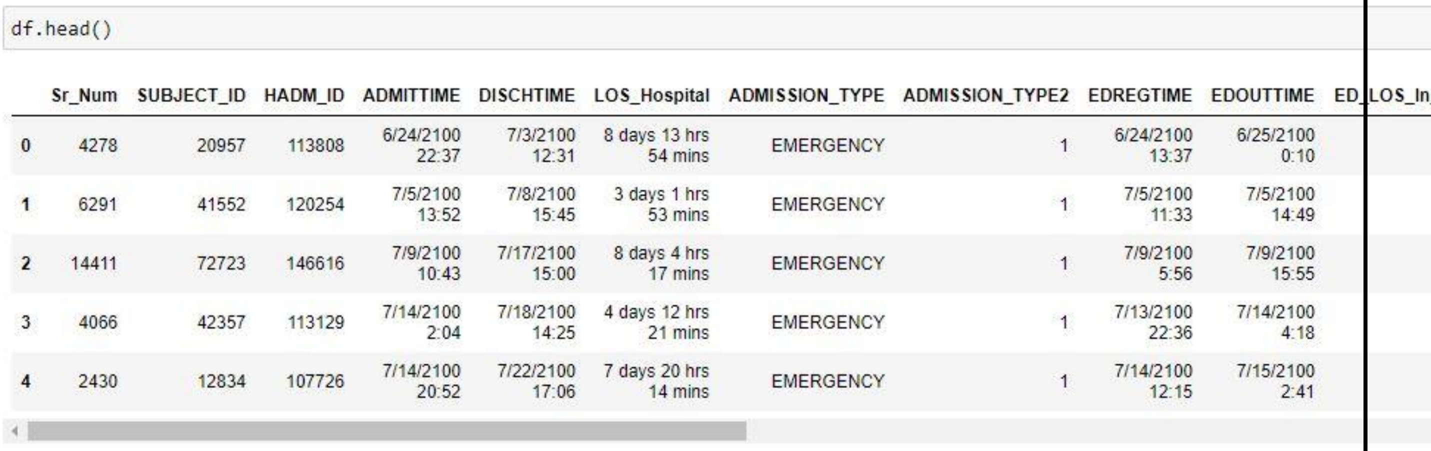


```
df.isnull().any()
```

```
Sr_Num                 False
SUBJECT_ID             False
HADM_ID                False
ADMITTIME              False
DISCHTIME              False
LOS_Hospital           False
ADMISSION_TYPE         False
ADMISSION_TYPE2        False
EDREGTIME              False
EDOUTTIME              False
ED_LOS_In_Seconds      False
ED_LOS_In_Mins         False
LOS_EmergencyDept      False
DIAGNOSIS              False
DIAGNOSIS_CODE         False
HOSPITAL_EXPIRE_FLAG   False
ICD9_CODES             False
dtype: bool
```

None of the column contains NULL value

Data visualization in Python:



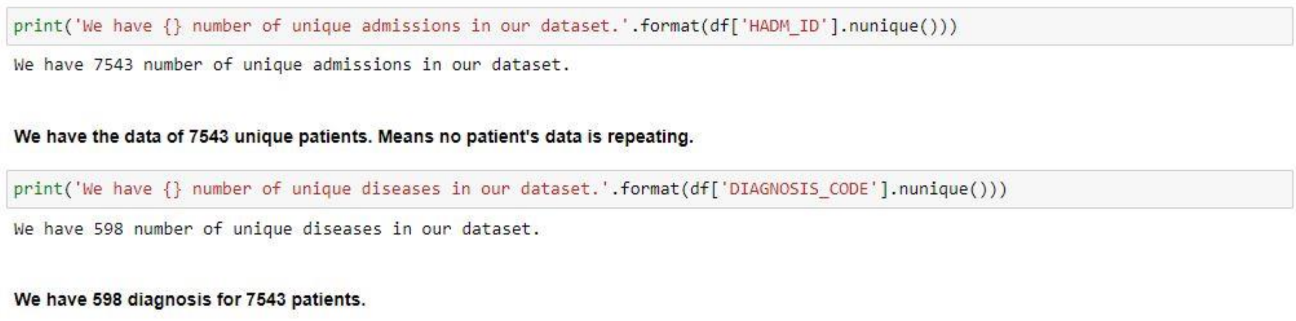
```
df.head()
```

	Sr_Num	SUBJECT_ID	HADM_ID	ADMITTIME	DISCHTIME	LOS_Hospital	ADMISSION_TYPE	ADMISSION_TYPE2	EDREGTIME	EDOUTTIME	ED_LOS_In
0	4278	20957	113808	6/24/2100 22:37	7/3/2100 12:31	8 days 13 hrs 54 mins	EMERGENCY	1	6/24/2100 13:37	6/25/2100 0:10	
1	6291	41552	120254	7/5/2100 13:52	7/8/2100 15:45	3 days 1 hrs 53 mins	EMERGENCY	1	7/5/2100 11:33	7/5/2100 14:49	
2	14411	72723	146616	7/9/2100 10:43	7/17/2100 15:00	8 days 4 hrs 17 mins	EMERGENCY	1	7/9/2100 5:56	7/9/2100 15:55	
3	4066	42357	113129	7/14/2100 2:04	7/18/2100 14:25	4 days 12 hrs 21 mins	EMERGENCY	1	7/13/2100 22:36	7/14/2100 4:18	
4	2430	12834	107726	7/14/2100 20:52	7/22/2100 17:06	7 days 20 hrs 14 mins	EMERGENCY	1	7/14/2100 12:15	7/15/2100 2:41	

Our target variable will be LOS(in minutes), And we will predict LOS on the basis of diagnosis code

Overview of our data. How our data is placed in dataset.

Patients and diagnosis data:



```
print('We have {} number of unique admissions in our dataset.'.format(df['HADM_ID'].nunique()))
```

```
We have 7543 number of unique admissions in our dataset.
```

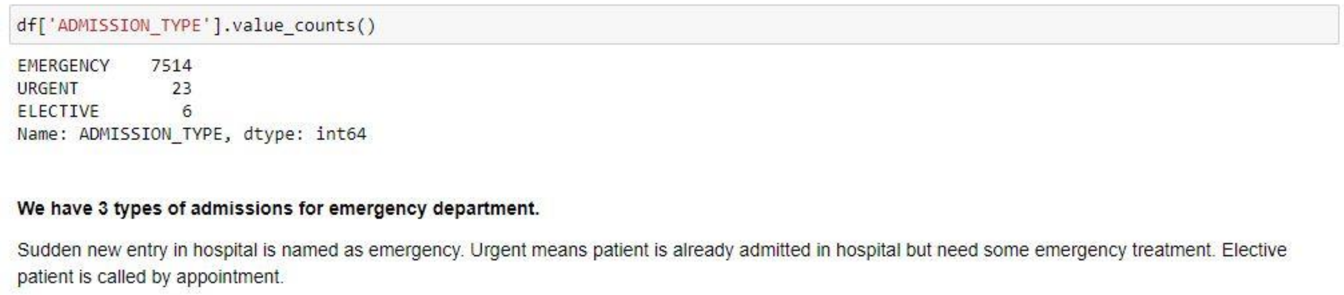
We have the data of 7543 unique patients. Means no patient's data is repeating.

```
print('We have {} number of unique diseases in our dataset.'.format(df['DIAGNOSIS_CODE'].nunique()))
```

```
We have 598 number of unique diseases in our dataset.
```

We have 598 diagnosis for 7543 patients.

Information about emergency entries:



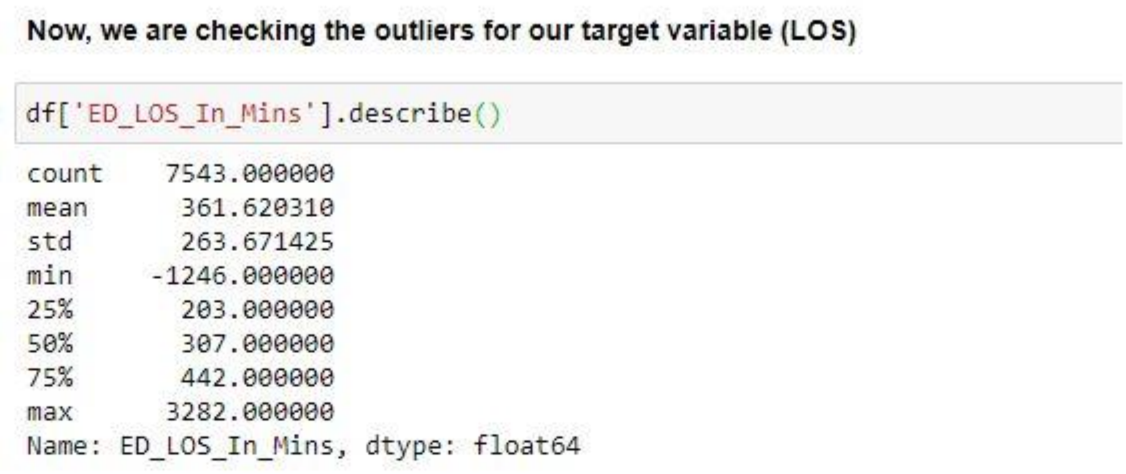
```
df['ADMISSION_TYPE'].value_counts()
```

```
EMERGENCY    7514
URGENT         23
ELECTIVE        6
Name: ADMISSION_TYPE, dtype: int64
```

We have 3 types of admissions for emergency department.

Sudden new entry in hospital is named as emergency. Urgent means patient is already admitted in hospital but need some emergency treatment. Elective patient is called by appointment.

Checking outliers:

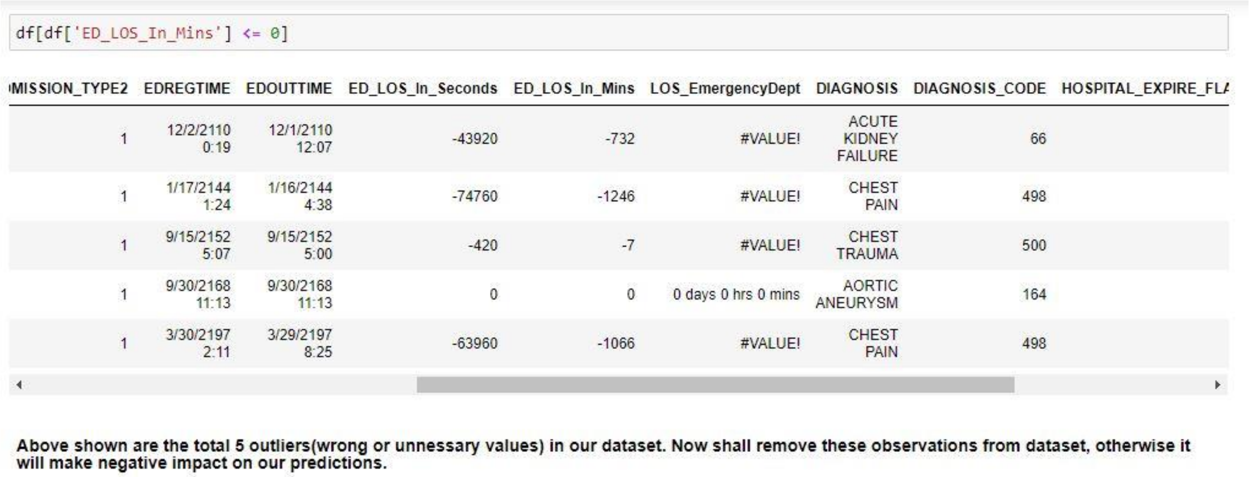


Now, we are checking the outliers for our target variable (LOS)

```
df['ED_LOS_In_Mins'].describe()
```

```
count    7543.000000
mean      361.620310
std       263.671425
min     -1246.000000
25%       203.000000
50%       307.000000
75%       442.000000
max      3282.000000
Name: ED_LOS_In_Mins, dtype: float64
```

Visualizing outliers:

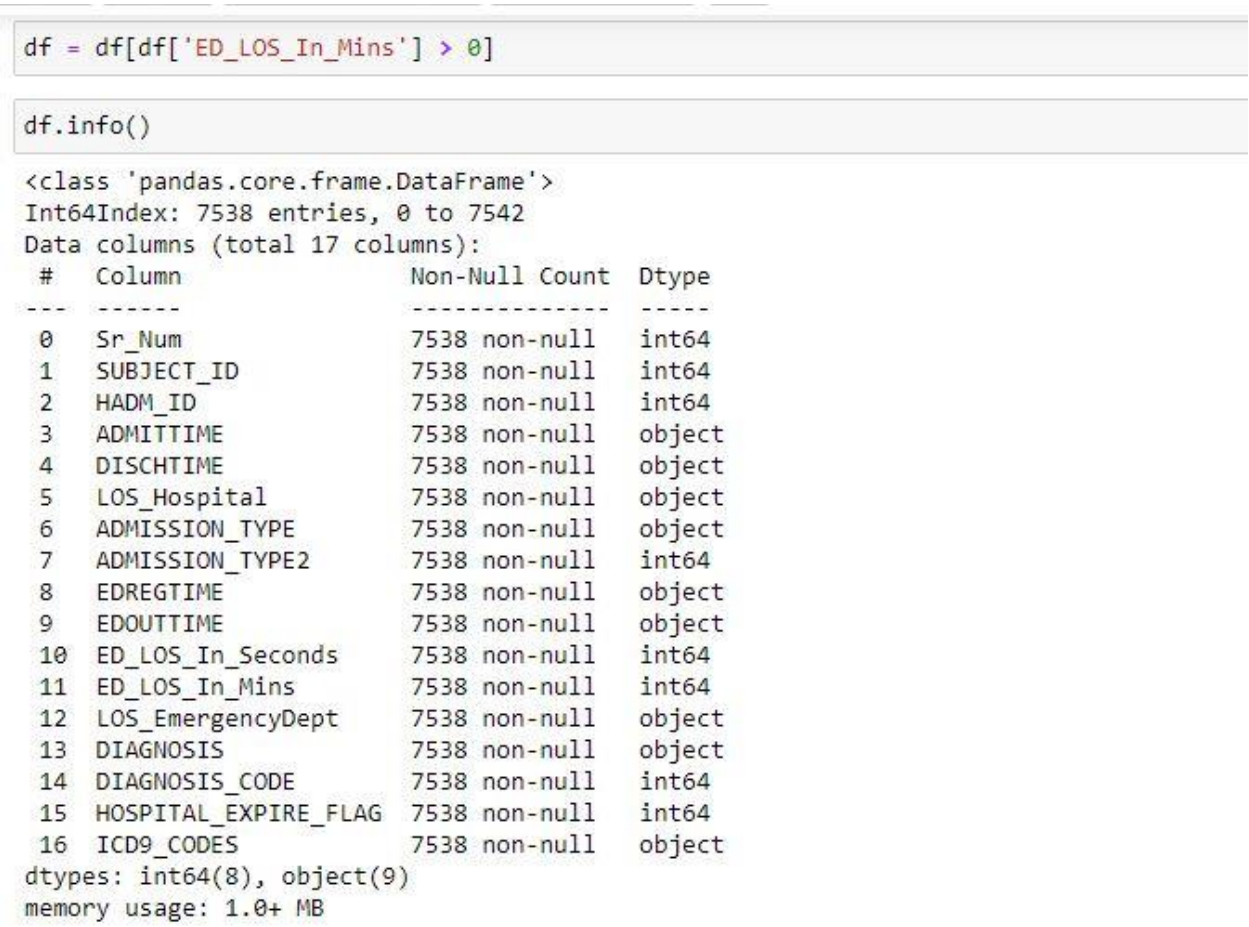


```
df[df['ED_LOS_In_Mins'] <= 0]
```

MISSION_TYPE2	EDREGTIME	EDOUTTIME	ED_LOS_In_Seconds	ED_LOS_In_Mins	LOS_EmergencyDept	DIAGNOSIS	DIAGNOSIS_CODE	HOSPITAL_EXPIRE_FL
1	12/2/2110 0:19	12/1/2110 12:07	-43920	-732	#VALUE!	ACUTE KIDNEY FAILURE	66	
1	1/17/2144 1:24	1/16/2144 4:38	-74760	-1246	#VALUE!	CHEST PAIN	498	
1	9/15/2152 5:07	9/15/2152 5:00	-420	-7	#VALUE!	CHEST TRAUMA	500	
1	9/30/2168 11:13	9/30/2168 11:13	0	0	0 days 0 hrs 0 mins	AORTIC ANEURYSM	164	
1	3/30/2197 2:11	3/29/2197 8:25	-63960	-1066	#VALUE!	CHEST PAIN	498	

Above shown are the total 5 outliers(wrong or unnessary values) in our dataset. Now shall remove these observations from dataset, otherwise it will make negative impact on our predictions.

Removing outliers:

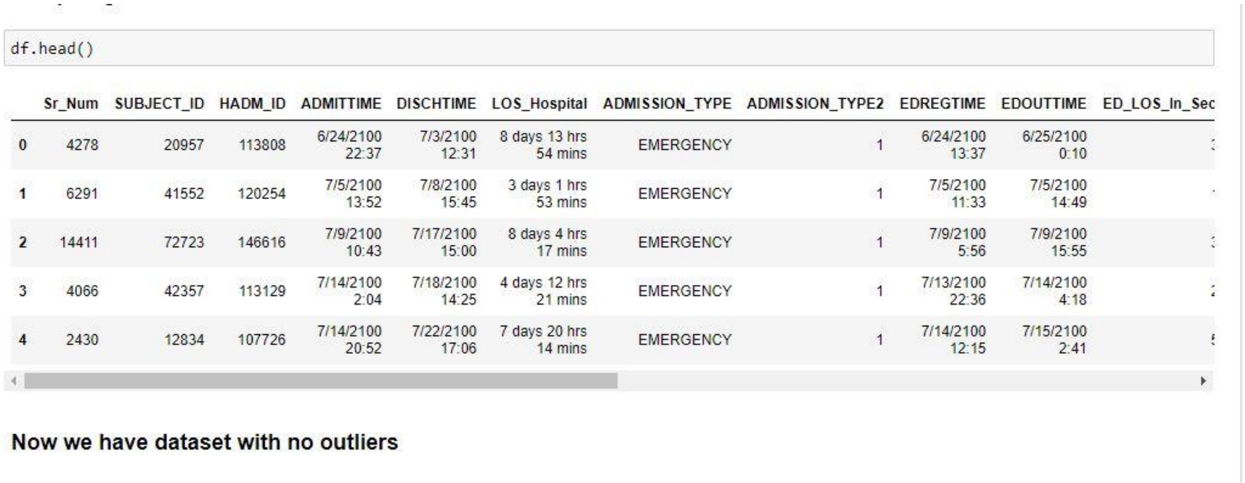


```
df = df[df['ED_LOS_In_Mins'] > 0]
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7538 entries, 0 to 7542
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Sr_Num              7538 non-null   int64
 1   SUBJECT_ID          7538 non-null   int64
 2   HADM_ID             7538 non-null   int64
 3   ADMITTIME           7538 non-null   object
 4   DISCHTIME           7538 non-null   object
 5   LOS_Hospital        7538 non-null   object
 6   ADMISSION_TYPE      7538 non-null   object
 7   ADMISSION_TYPE2     7538 non-null   int64
 8   EDREGTIME           7538 non-null   object
 9   EDOUTTIME           7538 non-null   object
 10  ED_LOS_In_Seconds   7538 non-null   int64
 11  ED_LOS_In_Mins      7538 non-null   int64
 12  LOS_EmergencyDept   7538 non-null   object
 13  DIAGNOSIS           7538 non-null   object
 14  DIAGNOSIS_CODE      7538 non-null   int64
 15  HOSPITAL_EXPIRE_FLAG  7538 non-null   int64
 16  ICD9_CODES          7538 non-null   object
dtypes: int64(8), object(9)
memory usage: 1.0+ MB
```

Data after removing outliers:

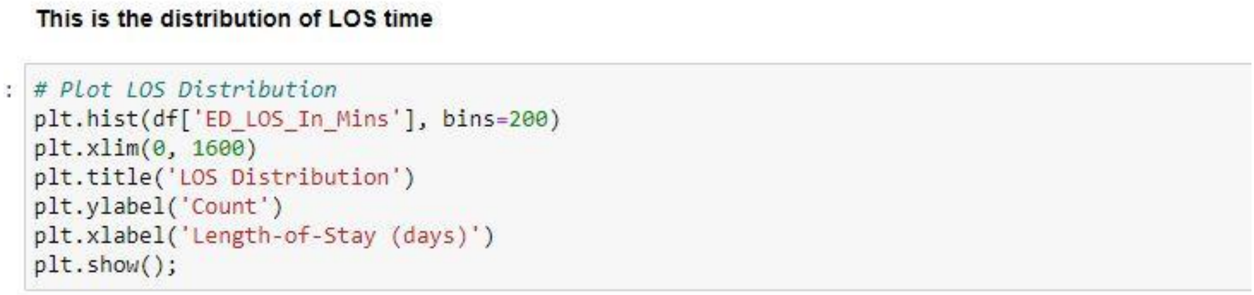


```
df.head()
```

	Sr_Num	SUBJECT_ID	HADM_ID	ADMITTIME	DISCHTIME	LOS_Hospital	ADMISSION_TYPE	ADMISSION_TYPE2	EDREGTIME	EDOUTTIME	ED_LOS_In_Sec
0	4278	20957	113808	6/24/2100 22:37	7/3/2100 12:31	8 days 13 hrs 54 mins	EMERGENCY	1	6/24/2100 13:37	6/25/2100 0:10	
1	6291	41552	120254	7/5/2100 13:52	7/8/2100 15:45	3 days 1 hrs 53 mins	EMERGENCY	1	7/5/2100 11:33	7/5/2100 14:49	
2	14411	72723	146616	7/9/2100 10:43	7/17/2100 15:00	8 days 4 hrs 17 mins	EMERGENCY	1	7/9/2100 5:56	7/9/2100 15:55	
3	4066	42357	113129	7/14/2100 2:04	7/18/2100 14:25	4 days 12 hrs 21 mins	EMERGENCY	1	7/13/2100 22:36	7/14/2100 4:18	
4	2430	12834	107726	7/14/2100 20:52	7/22/2100 17:06	7 days 20 hrs 14 mins	EMERGENCY	1	7/14/2100 12:15	7/15/2100 2:41	

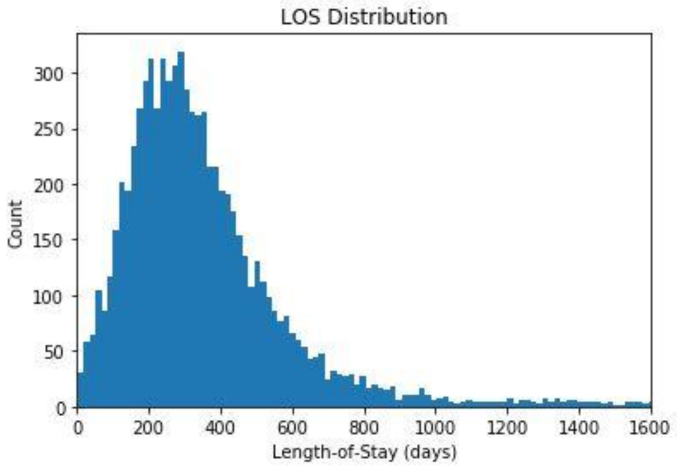
Now we have dataset with no outliers

Distribution of LOS:



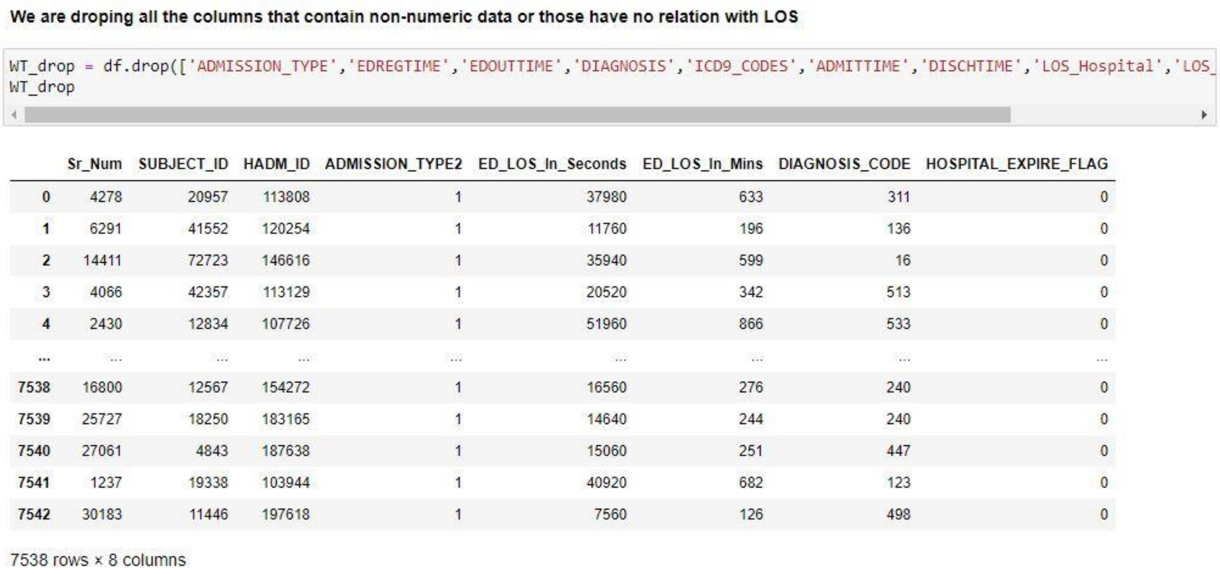
This is the distribution of LOS time

```
# Plot LOS Distribution
plt.hist(df['ED_LOS_In_Mins'], bins=200)
plt.xlim(0, 1600)
plt.title('LOS Distribution')
plt.ylabel('Count')
plt.xlabel('Length-of-Stay (days)')
plt.show();
```



Although it is right skewed, but its bell shaped so, we don’t need scaling.

Preparation for the training and testing data:



We are droping all the columns that contain non-numeric data or those have no relation with LOS

```
WT_drop = df.drop(['ADMISSION_TYPE','EDREGTIME','EDOUTTIME','DIAGNOSIS','ICD9_CODES','ADMITTIME','DISCHTIME','LOS_Hospital','LOS_
WT_drop
```

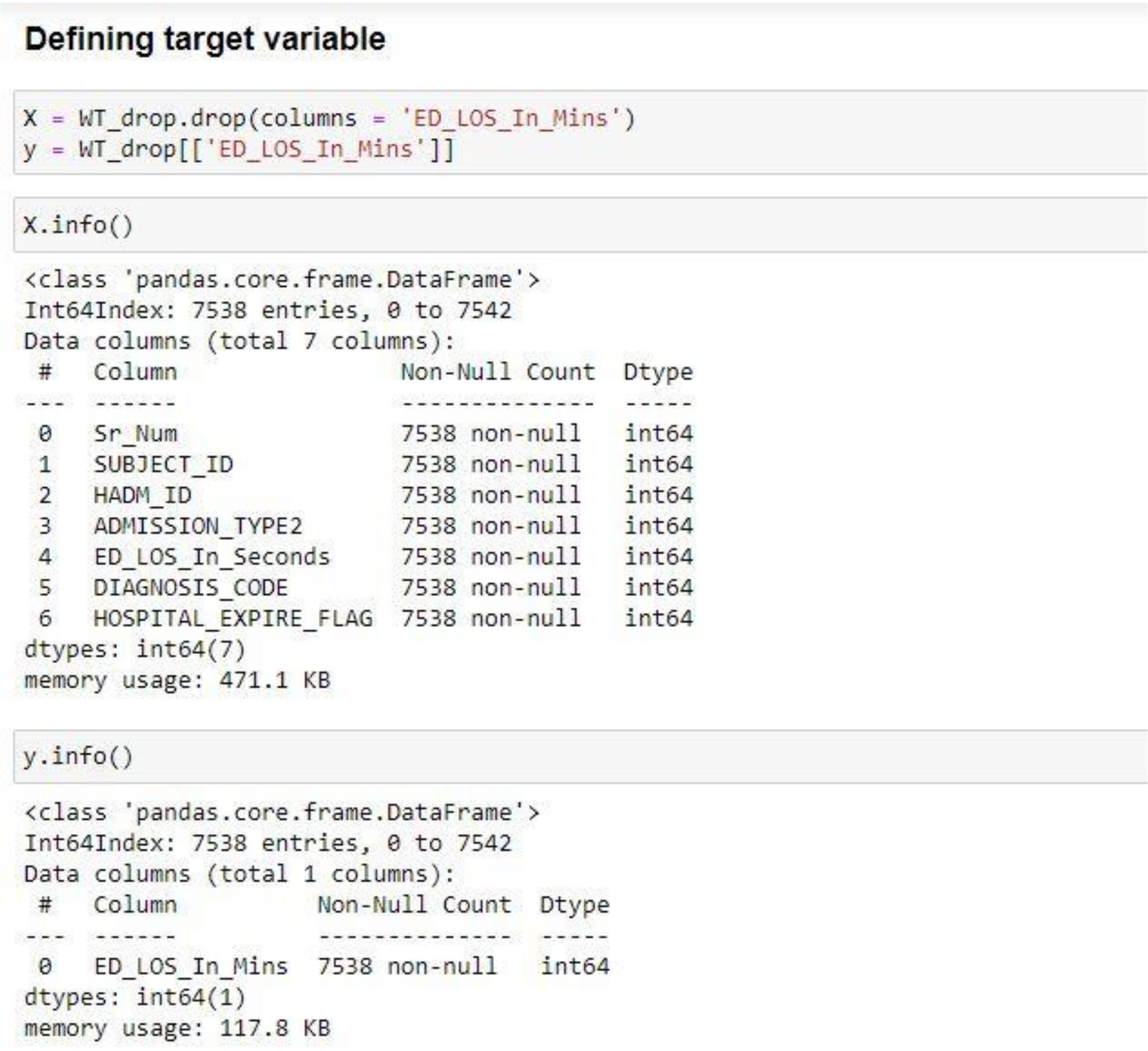
	Sr_Num	SUBJECT_ID	HADM_ID	ADMISSION_TYPE2	ED_LOS_In_Seconds	ED_LOS_In_Mins	DIAGNOSIS_CODE	HOSPITAL_EXPIRE_FLAG
0	4278	20957	113808	1	37980	633	311	0
1	6291	41552	120254	1	11760	196	136	0
2	14411	72723	146616	1	35940	599	16	0
3	4066	42357	113129	1	20520	342	513	0
4	2430	12834	107726	1	51960	866	533	0
...	...	...	...	...	...	...	...	...
7538	16800	12567	154272	1	16560	276	240	0
7539	25727	18250	183165	1	14640	244	240	0
7540	27061	4843	187638	1	15060	251	447	0
7541	1237	19338	103944	1	40920	682	123	0
7542	30183	11446	197618	1	7560	126	498	0

7538 rows × 8 columns

We have removed all the non-numeric variables as our prediction is a continues numeric value so, we are using regression techniques.

Defining target variable:

From all the numeric variables we are dropping length of stay variable because our prediction is based on it. So, we are assuming it on Y axis and others on X axis.



Defining target variable

```
X = WT_drop.drop(columns = 'ED_LOS_In_Mins')
y = WT_drop[['ED_LOS_In_Mins']]
```

```
X.info()
```

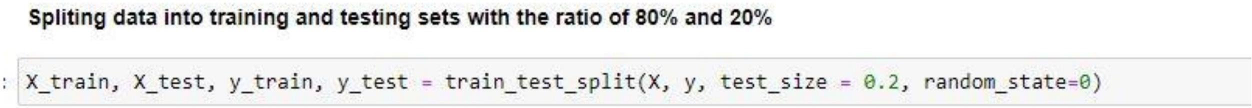
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7538 entries, 0 to 7542
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Sr_Num                7538 non-null   int64
 1   SUBJECT_ID            7538 non-null   int64
 2   HADM_ID               7538 non-null   int64
 3   ADMISSION_TYPE2       7538 non-null   int64
 4   ED_LOS_In_Seconds     7538 non-null   int64
 5   DIAGNOSIS_CODE        7538 non-null   int64
 6   HOSPITAL_EXPIRE_FLAG  7538 non-null   int64
dtypes: int64(7)
memory usage: 471.1 KB
```

```
y.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7538 entries, 0 to 7542
Data columns (total 1 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   ED_LOS_In_Mins  7538 non-null   int64
dtypes: int64(1)
memory usage: 117.8 KB
```

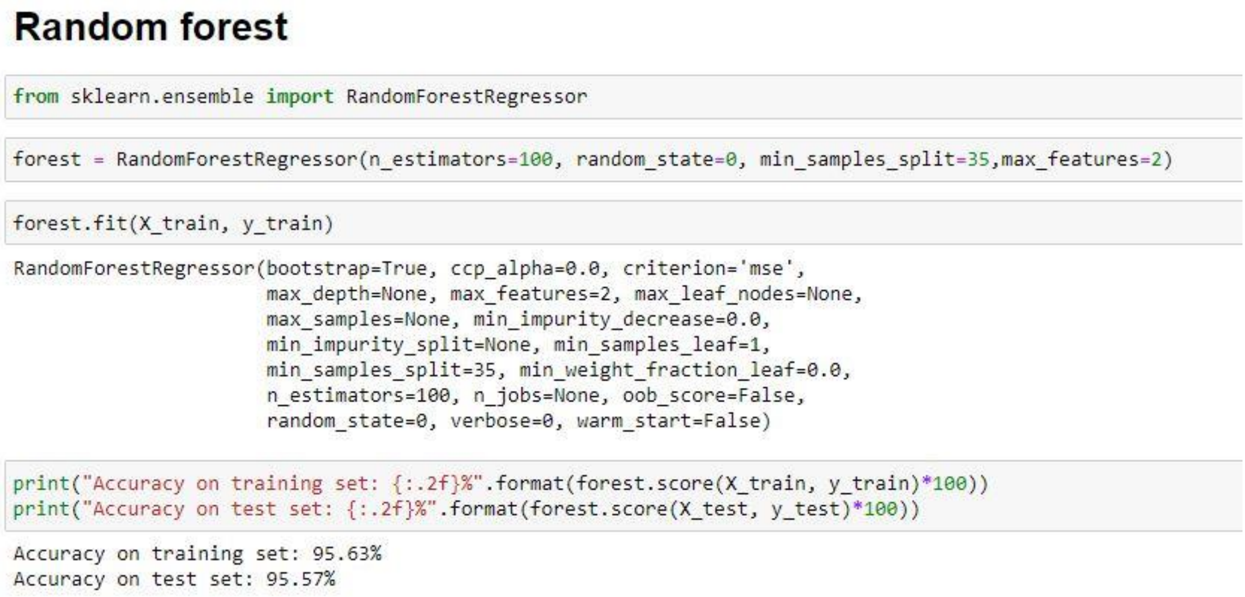
Model Building

Splitting data:



Spliting data into training and testing sets with the ratio of 80% and 20%

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state=0)
```



Random forest

```
from sklearn.ensemble import RandomForestRegressor
```

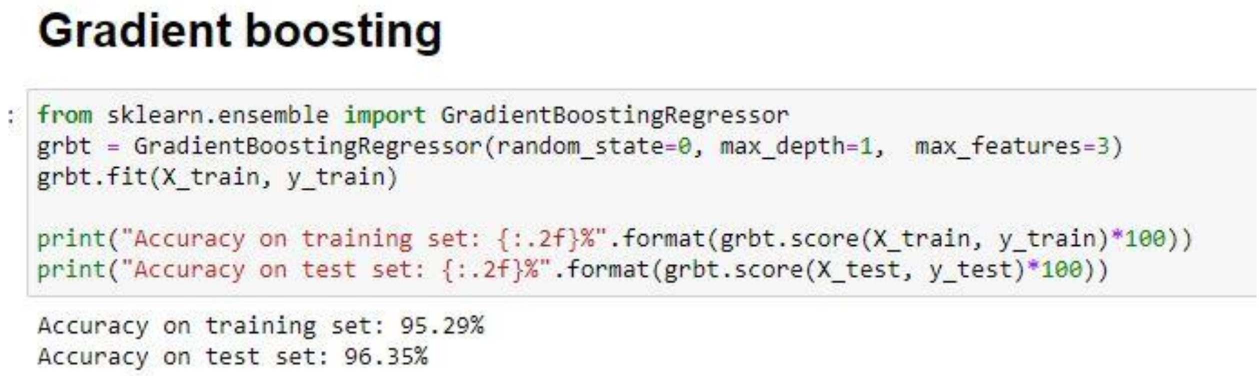
```
forest = RandomForestRegressor(n_estimators=100, random_state=0, min_samples_split=35,max_features=2)
```

```
forest.fit(X_train, y_train)
```

```
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                      max_depth=None, max_features=2, max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=35, min_weight_fraction_leaf=0.0,
                      n_estimators=100, n_jobs=None, oob_score=False,
                      random_state=0, verbose=0, warm_start=False)
```

```
print("Accuracy on training set: {:.2f}%".format(forest.score(X_train, y_train)*100))
print("Accuracy on test set: {:.2f}%".format(forest.score(X_test, y_test)*100))
```

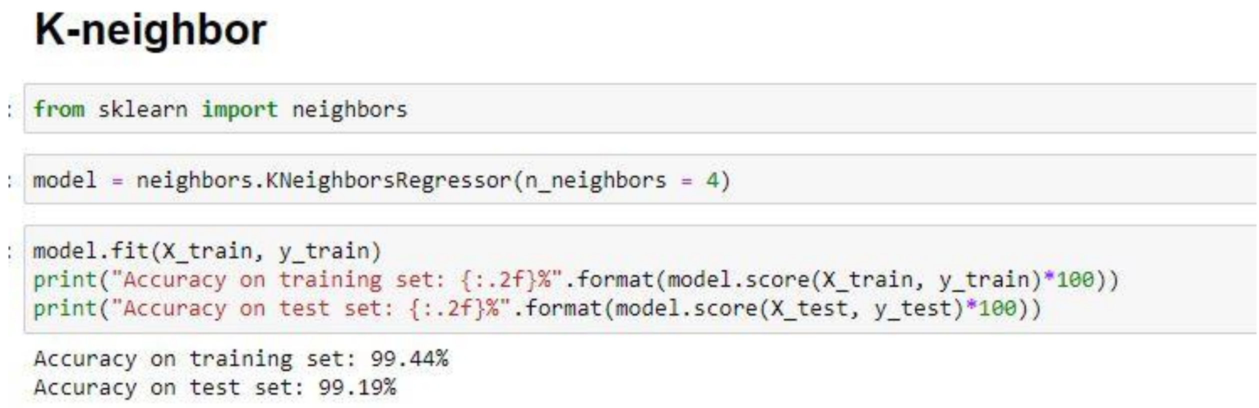
```
Accuracy on training set: 95.63%
Accuracy on test set: 95.57%
```



Gradient boosting

```
from sklearn.ensemble import GradientBoostingRegressor
grbt = GradientBoostingRegressor(random_state=0, max_depth=1,  max_features=3)
grbt.fit(X_train, y_train)

print("Accuracy on training set: {:.2f}%".format(grbt.score(X_train, y_train)*100))
print("Accuracy on test set: {:.2f}%".format(grbt.score(X_test, y_test)*100))
```

```
Accuracy on training set: 95.29%
Accuracy on test set: 96.35%
```



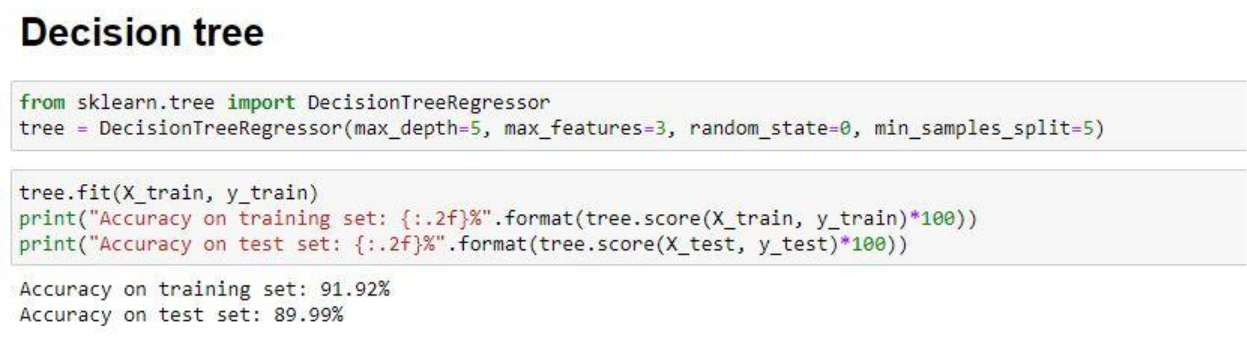
K-neighbor

```
from sklearn import neighbors
```

```
model = neighbors.KNeighborsRegressor(n_neighbors = 4)
```

```
model.fit(X_train, y_train)
print("Accuracy on training set: {:.2f}%".format(model.score(X_train, y_train)*100))
print("Accuracy on test set: {:.2f}%".format(model.score(X_test, y_test)*100))
```

```
Accuracy on training set: 99.44%
Accuracy on test set: 99.19%
```

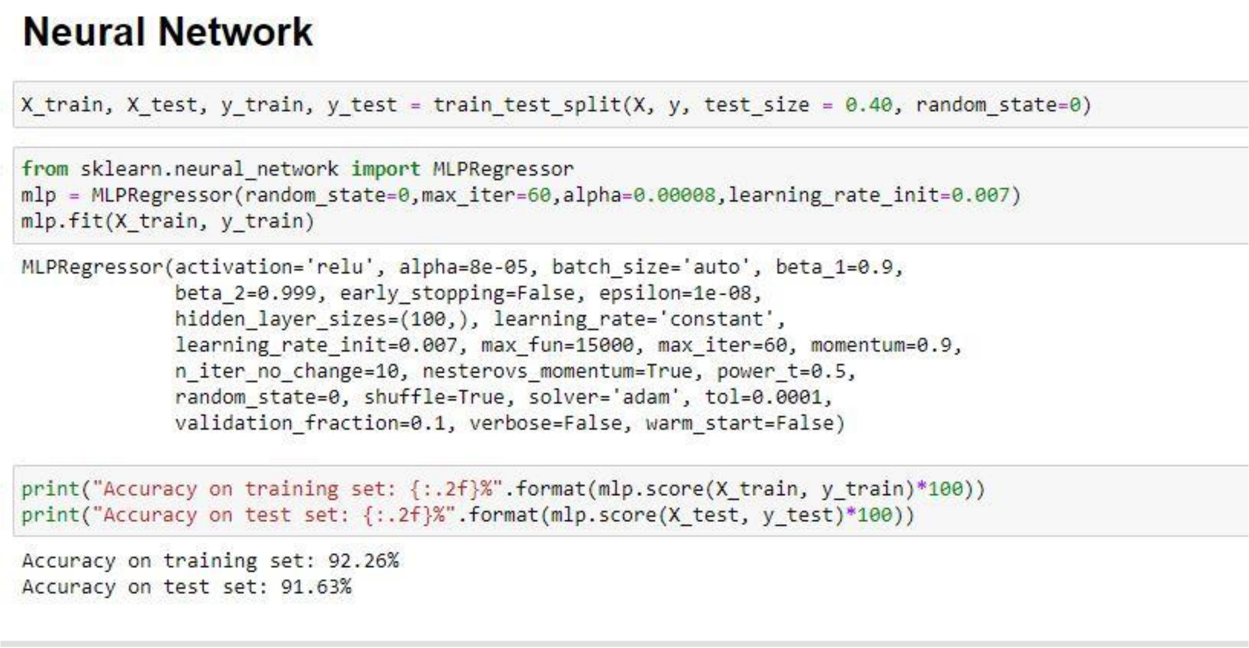


Decision tree

```
from sklearn.tree import DecisionTreeRegressor
tree = DecisionTreeRegressor(max_depth=5, max_features=3, random_state=0, min_samples_split=5)
```

```
tree.fit(X_train, y_train)
print("Accuracy on training set: {:.2f}%".format(tree.score(X_train, y_train)*100))
print("Accuracy on test set: {:.2f}%".format(tree.score(X_test, y_test)*100))
```

```
Accuracy on training set: 91.92%
Accuracy on test set: 89.99%
```



Neural Network

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.40, random_state=0)
```

```
from sklearn.neural_network import MLPRegressor
mlp = MLPRegressor(random_state=0,max_iter=60,alpha=0.00008,learning_rate_init=0.007)
mlp.fit(X_train, y_train)
```

```
MLPRegressor(activation='relu', alpha=8e-05, batch_size='auto', beta_1=0.9,
             beta_2=0.999, early_stopping=False, epsilon=1e-08,
             hidden_layer_sizes=(100,), learning_rate='constant',
             learning_rate_init=0.007, max_fun=15000, max_iter=60, momentum=0.9,
             n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5,
             random_state=0, shuffle=True, solver='adam', tol=0.0001,
             validation_fraction=0.1, verbose=False, warm_start=False)
```

```
print("Accuracy on training set: {:.2f}%".format(mlp.score(X_train, y_train)*100))
print("Accuracy on test set: {:.2f}%".format(mlp.score(X_test, y_test)*100))
```

```
Accuracy on training set: 92.26%
Accuracy on test set: 91.63%
```

Conclusion

In this project, we have noticed that the patients who need under given diagnosis should be taken seriously and given priority:

1). Altered Mental Status
2). Abdominal Pain
3). Cardiac Arrest
4). Chest Pain
5). Acute Kidney Failure

And to find the best LOS, the Neural network model trained by us can be used. Because as compare to other models neural network gives us very best accuracy. Neither it underestimates nor overestimates.

Goals

- Predicting best LOS at preadmission stage
- Saving lives by taking patients on priority basis
- Reduces patients' unnecessary expenses by reducing LOS
- Reducing waiting time

Challenges

- Dataset, as we are beginners and this project was new for us so, it was very difficult find this large and reliable dataset.
- Data cleaning, ICD codes were containing alphanumeric characters, but our requirement was only numeric codes. So, we needed to do in ourselve for all diagnosis. It was very time consuming.
- Modelling, understanding neural networks and implementation was a big challenge also.
- Controlling accuracies, all the models were overpredicting the accuracies. When we were playing with combination of all the parameters of models' functions the accuracies were starting fluctuating (sometimes very underfitting or sometimes overfitting). We tried many combinations of data splitting ratio but none of that was working. At the end the accuracies our models are giving are best from our side.
- Due to COVID-19 for social distancing we were working with the help of Github, Whatsapp, Emails, Phone calls.

Contribution

For the good management and flow of project, we splitted our project work into major tasks and further sub tasks. To keep track of each task and project progress, we were using a contribution chart and updating it at the completion of each task. That chart is:

Major Task	**Member Name**	**Sub Tasks**	**Description**	**Completed on**	**Status**
Topic selection	Kanchan Bagga	Finding project topic, Topic related search	Project that could be useful in future. Should have something new. Scope for job placement. Should be useful in real life as a tool.	February 04, 2020	Done
Dataset	Jaspreet Kaur	Finding dataset (Searching, getting access and downloading)	Searched dataset from different websites. Sent emails to many of the organizations for accessing their datasets. Got access for MIMIC dataset and downloaded	February 18, 2020	Done
	Jaspreet Kaur Kanchan Bagga Varinderjit Singh	Target and supporting variables selection	Target variable (LOS) was already cleared in everyone's mind but all the supporting variables were needed to select that could effect on LOS.	February 25, 2020	Done
	Kanchan Bagga	Data Cleaning (Changing format, keeping necessary data only) using excel	Changed downloaded files format. Made one single file from 27 files by merging all necessary columns in master file only. Removed unnecessary columns only, removed outliers, data about emergency department only was kept. Some new columns are made. Missing data rows are removed.	March 03, 2020	Done
	Varinderjit Singh	Cleaned data understanding & exploration	Understood & explored cleaned data in SQL Server for my satisfaction and cross checking.	March 05, 2020	Done
Project related research	Jaspreet Kaur	Searching work done on related projects	Projects that are already made on same topic. So that we could come to know what is done and what different we can do.	March 03,2020	Done

Article reading	Varinderjit Singh Kanchan Bagga Jaspreet Kaur	Studied 10 (4-3-3) articles individually	Study articles based on how we can use machine learning in our project. For understanding what kind of models, we can use and how we can improve their accuracy.	March 12, 2020	Done
Model Study	Jaspreet Kaur	ANN, Logistic regression, K-neighbor	YouTube tutorials, Google search	March 23, 2020	Done
	Kanchan Bagga	CNN, Random forest, Gradient boosting	YouTube tutorials, Google search	March 19, 2020	Done
	Varinderjit Singh	Confusion matrix, Multiple linear regression	YouTube tutorials, Google search	March 20, 2020	Done
Model Building	Varinderjit Singh	Neural network, Multiple linear regression, Confusion matrix		March 19, 2020	Done
	Kanchan Bagga	Random forest, Gradient boosting		March 24, 2020	Done
	Jaspreet Kaur	Logistic regression, K-neighbor		March 24, 2020	Done
Report writing	Jaspreet Kaur			April 17, 2020	Done
Presentation	Kanchan Bagga			April 9, 2020	Done

References

- https://physionet.org/content/mimiciii/1.4/
- https://www.oecd-ilibrary.org/docserver/health_glance-2017-64-en.pdf?expires=1587195689&id=id&accname=guest&checksum=57D762A78EB8089272B9AC847B5E4EA0
- http://www.gov.pe.ca/photos/original/src_leanpfhsh.pdf
- https://www.healthcatalyst.com/success_stories/reducing-length-of-stay-in-hospital
- https://youtu.be/BhpvH5DuVu8
- https://www.youtube.com/watch?v=GvQwE2OhL8I
- https://www.youtube.com/watch?v=aircAruvnKk