**Project of Capstone**

**On**

**Length of Hospital Stay Prediction**

**St. Clair College, Windsor, ON**

**Submitted to**

Profs. Muhammad Shahid

**Submitted By**

|  |  |
| --- | --- |
| NAMES | STUDENT ID |
| Jaspreet Kaur | 0730470 |
| Kanchan Bagga | 0732356 |
| Varinderjit Singh | 0730482 |

**Date- 04/17/2020**

**Final Project Report**

**DAB 402**

**CAPSTONE PROJECT**

**Content:**

* **Topic (Problem Statement)**
* **Literature Review**
* **Ethical Concerns**
* **Dataset**
* **Methodology**
* Data Overview
* How the dataset fit the problem
* Cleaning and Preprocessing
* Balancing
* Scaling
* Modeling
* **Conclusion**
* **Goals**
* **Challenges**
* **Results**
* **Contributions**
* **References**

**Title**: Length of Hospital Stay Prediction

**Problem Statement**

**Introduction:**

Today, Hospitals are faced the main problem regarding holding the patients to the cure mainly in the emergency department because general patients and serious patients have an equal waiting time for the treatment to cure with the physical sickness or any disorder, which is very big problem. To eliminating the length of stay in hospital we are using the different techniques for prediction about this big issue to help the hospitals. Predicting the length of stay improve the hospital planning and manage the resources used for patient’s 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 any rate $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. This topic is not our first priority or selection to work on, initially, we planned to write a dataset on the basis of nature that topic was natural disasters and how these disasters impact on the environment’s own balance as well as the economy of the country and the native public or citizens of the belonging nation. Our main interest about that topic was the big natural issue in Australia and the harmful effects over there. But due to the lack of dataset and general information regarding that so we leave that topic and move to the dataset of health care field. However, we easily got every basic information and dataset related to this field, which further proven very helpful to us in order to complete the dataset on the field of health care. 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 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 quick as it is possible. Because there are various measures and models are calculated to ensure the real and accurate 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, due to their nature of nonlinear statistical principles. 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.

**Methodology**

In this project we were used the different model techniques to get the better accuracy to make the prediction because at the end we find the best model to get the best accuracy to predict the length of stay in hospitals for general patients in emergency departments.

There are different models:

**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 is also work on the classification and regression techniques. The random forest is a classification algorithm containing of numerous decision trees. 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 random forest is the regressor. 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 control the overfitting. But when we apply on our dataset it shows also 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 form 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.

**Reference**

**References:**

* Rosebrock, A. (2020). Machine Learning in Python - PyImageSearch. Retrieved 18 April 2020, from <https://www.pyimagesearch.com/2019/01/14/machine-learning-in-python/>
* (2020). Retrieved 18 April 2020, from <https://adventuresinmachinelearning.com/wp-content/uploads/2017/07/An-introduction-to-neural-networks-for-beginners.pdf>
* Multilayer perceptron. (2020). Retrieved 18 April 2020, from <https://en.wikipedia.org/wiki/Multilayer_perceptron>
* Retrieved 17 April 2020, from <https://www.researchgate.net/publication/324177552_Predicting_Hospital_Length_of_stay>
* 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
* <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0201016>
* Retrieved 18 April 2020, from <https://towardsdatascience.com/predicting-hospital-length-of-stay-at-time-of-admission55dfdfe69598>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5898738/>
* 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/
* (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>
* 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.
* 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).

# (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