Logo, company name

Description automatically generated

**HAP 780**

**Data Mining in Health Care**

**Length of Stay of Patients Who are diagnosed**

**with Pneumonia**

**Abstract:**

**Objective:** In this study, we predict the Length of stay of patients in which we are predicting the duration of the patient being hospitalized from date of admission to date of discharge. We are mainly focusing on Length of stay of patients who are diagnosed with “PNEUMONIA”. Also, does the diagnosed patient stay more than 6 days or lesser than that can be known by creating various machine learning models.

**Data:** The data is taken from the website <https://www.physionet.org/>. In order to retrieve data, one has to undergo a training course from <https://www.citiprogram.org/index.cfm?pageID=14> by which there is a compliance training course to be completed. In the dataset, there are records of patients whose age is above 18.

**Method:** The tools we used for building the machine learning model is WEKA. The data is split into training and test data in the proportion of 80:20 respectively. The models implemented here are Bayesian Network, Naïve Bayes, Logistic Regression, Random Forest.

**Result:**

By running various models like Random Forest, Random Tree, Naïve Bayes, Logistic Regression we can observe there is a highest accuracy of 99.7% for random forest and 99.9% accuracy for Bayes net with 60% as train and 40 as test. With 80% as train and 20% as test we can observe the value of recall as 1, and by considering factors like ROC, Weighted average precision there is an accuracy of 99.7%

**Keywords:** MIMIC3, Prediction, Machine Learning models, Length of Stay.

**Introduction:**

Good health is one of the important aspects in our life. Maintaining good health, hygiene, being productive at work is very important for our day-to-day life and we are able to achieve this with proper various factors and one of the key factors is improving healthcare. And with healthcare data mining is one of the important applications to find useful and knowledgeable patterns by analyzing large amounts of data. By using these data patterns, we are now able to understand the trends, how the health care is working, factors which can be improved, factors leading to good health care. Some of the main reasons which data mining can help in healthcare industry is by improving the efficiency and decreasing the costs there by helping in the improvement of patient life.

Data mining is used in various applications like customer experience and satisfaction, customer relation management, detection of fraud and abuse of medical claims, product safety, improvement in the quality and need of medicine there by helping the patients.

Data mining can help cut the costs of healthcare and help to drive the efficiency and quality of healthcare in the future. Also, some of the main problems which can be overcome in near future are getting the right medicine, detecting false claims by insurance providers, choosing the best treatment plan based on previous trends.

Length of stay of a patient can be defined as the duration of the patient where he is admitted in the hospital to the time the patient is discharged. Predicting the length of stay of patients helps the healthcare providers in recommending the right diagnosis, allocation of necessary equipment and bed, arranging the proper type of medicine and specialists required.

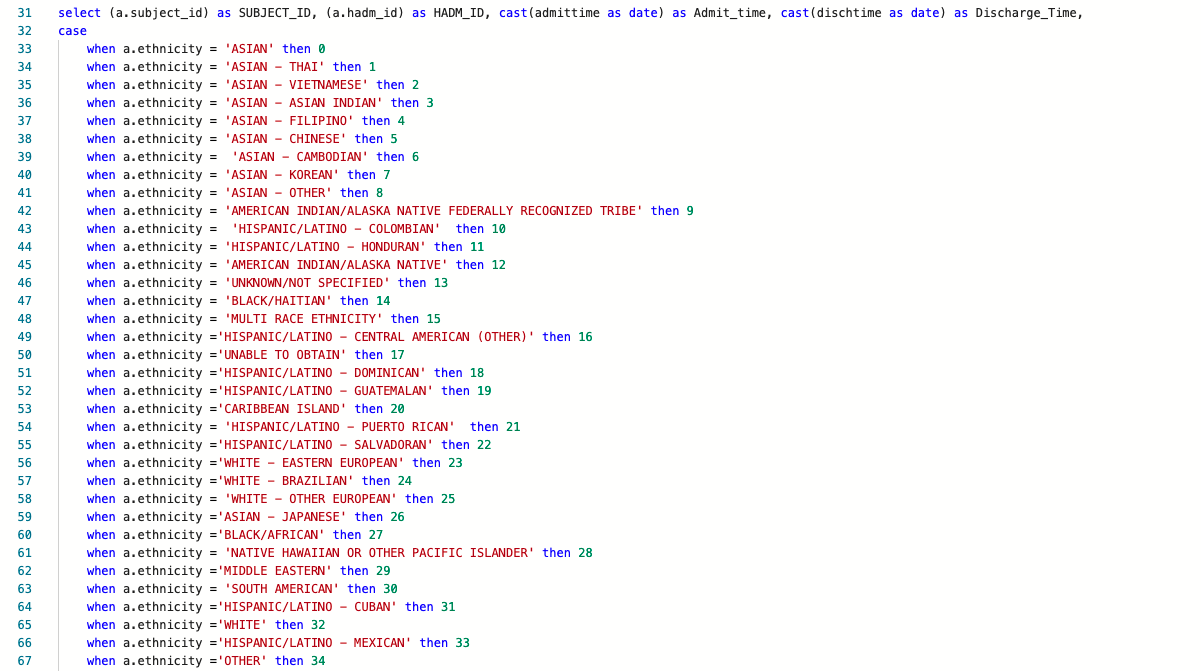
**Data Preprocessing:**

Data Preprocessing is most important and first step to do before building the machine learning model. Data might be in many forms and they need to be treated as the machines do not understand free text, images and videos. Preprocessing consists of treating the missing values, checking for Inconsistent values, checking for duplicate values.

The first step we are doing here is taking the Admissions table and by choosing the attributes like subject ID, Hospital ID, Admit time, Discharge Time, Ethnicity, Admission location, and type of insurance. We are then casting dischtime as date and admittime as date.

Data Pre-Processing has various steps:

First, I selected attributes like subject\_id, hadm\_id, admittime and discharge\_time and changed thnicity, admission\_location, admission\_type, and insurance to numeric values.

****

**Table

Description automatically generated**

****

**Graphical user interface, text, application

Description automatically generated**

Next, I changed the Gender to binary format.

A picture containing text

Description automatically generated

Changing Gender is important for us to calculate the length of stay and age.

Now, we are going to join the columns like gender and dob to calculate age and length of stay.

Graphical user interface, text, application, email

Description automatically generated

Here, the age and length of stay is calculated using attributes like discharge time, date of birth, admission time.

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

The patients who are dead should be deleted. So, we can there is some abnormality in the age column. Those are removed.

Graphical user interface, text, application

Description automatically generated

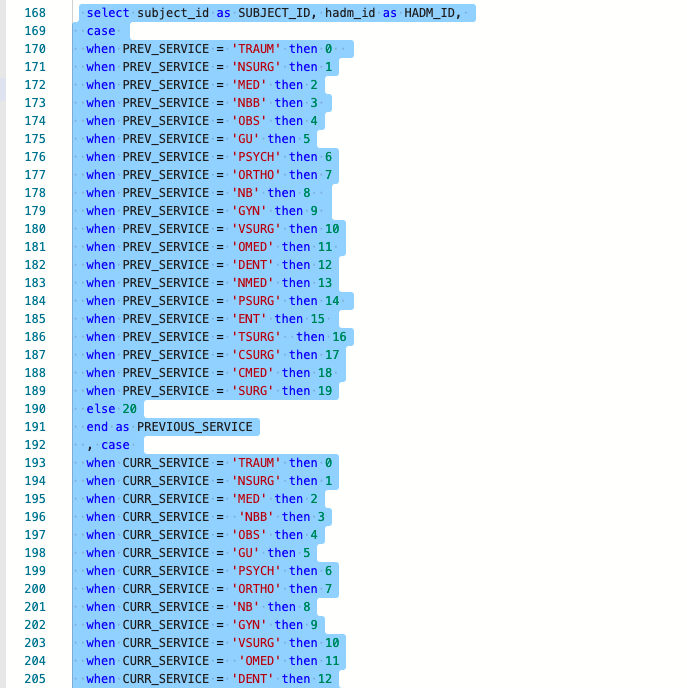
A picture containing text

Description automatically generated

Table

Description automatically generated

Next, we are assigning numeric values to attributes like prev\_service, and Curr\_service which are previous service and current service respectively.



A picture containing text

Description automatically generated

Graphical user interface, text, application

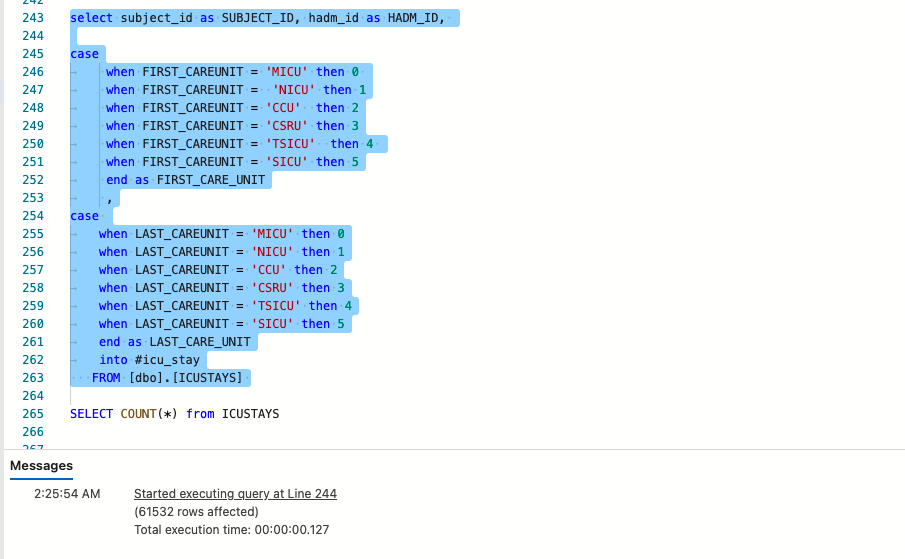
Description automatically generated

Next, the attributes like subject\_id, hadm\_id, CPT\_number are taken from table CPT Events and an average of CPT\_number is taken.

**Text

Description automatically generated**

Next, the columns which are needed from ICU stays table is selected and conversion takes place into numerical values.

****

**Table

Description automatically generated**

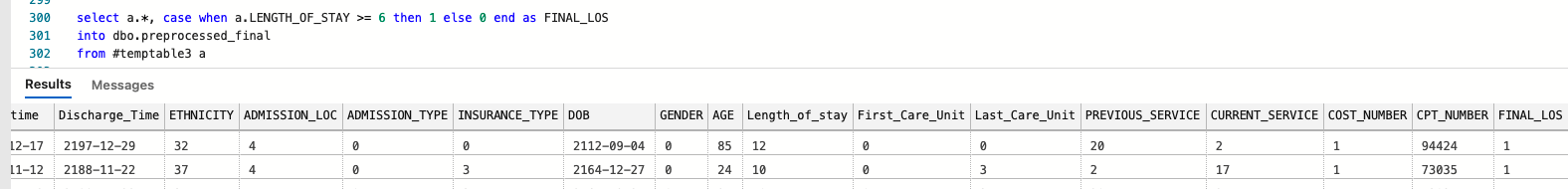
Next,the tables are joined, and we get the final preprocessed dataset for further analysis.

Graphical user interface, text, application, email

Description automatically generated

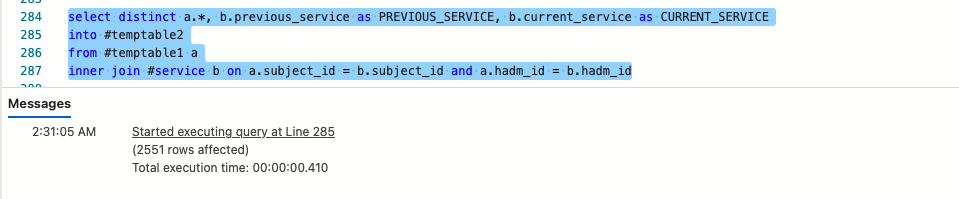
Table

Description automatically generated



Table

Description automatically generated

****

**Table

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Table

Description automatically generated**

Finally, the LOS is taken as 1 and 0 if the length of stay is 6 days and length of stay is more than 6 days respectively.

**Preprocessing in WEKA:**

Preprocessing has to be done in WEKA after uploading the csv file we get after preprocessing to remove unwanted attributes as we do not need them, and there has to be a conversion of attributes from Numeric to Nominal in order to visualize the data and perform analysis.

The columns that are removed are subject\_id, hadm\_id, admit\_time, discharge\_time, as we already used them to predict length of stay if they are more than 6 days or less than that.

**Graphical user interface, application

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

**Analysis:**

**Bayes Net:**

**Graphical user interface, text, application

Description automatically generated**

**Random Forest:**

**Graphical user interface, text, application, email

Description automatically generated**

**Naïve Bayes:**

**Graphical user interface, text

Description automatically generated**

**Random Tree:**

**Graphical user interface, text, application, email

Description automatically generated**

**Logistic Regression:**

**Graphical user interface, text, application

Description automatically generated**

**Conclusion:**

By performing analysis and running various models we can see that about 99.9% accuracy with 60 percent as training and remaining as test set, and random forest with highest accuracy of 99.7% and has better results when compared to others by considering the factors like Recall, ROC and Weighted Average Precision.

**Future Work:**

By considering other attributes, we can check how they are related and in the same way we can check the length of stay for each disease and this can help the insurance providers and hospitals to prepare themselves for better treatment in order to reduce the expenses, and get the necessary requirements well in advance.

**References:**

1. [**https://mimic.physionet.org/**](https://mimic.physionet.org/)
2. [**https://mimic.physionet.org/about/mimic/**](https://mimic.physionet.org/about/mimic/)
3. [**https://healthcareinamerica.us/how-data-mining-is-changing-health-care-27c1e9b3b372**](https://healthcareinamerica.us/how-data-mining-is-changing-health-care-27c1e9b3b372)
4. [**https://www.usfhealthonline.com/resources/key-concepts/data-mining-in-healthcare/**](https://www.usfhealthonline.com/resources/key-concepts/data-mining-in-healthcare/)
5. [**https://www.ahigherbranch.com/health/#:~:text=Why%20Is%20Health%20The%20Most,devote%20to%20each%20of%20them**](https://www.ahigherbranch.com/health/#:~:text=Why%20Is%20Health%20The%20Most,devote%20to%20each%20of%20them)**.**
6. [**https://towardsdatascience.com/predicting-hospital-length-of-stay-at-time-of-admission-55dfdfe69598**](https://towardsdatascience.com/predicting-hospital-length-of-stay-at-time-of-admission-55dfdfe69598)
7. [**https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825**](https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825)