

St. Clair College of Applied Arts and Technology

Dab 304 Healthcare Analytics

Title

Admission analysis on mimic III

Group member:

Preetkumar SHAH, 0759151

Mitanshi Patel, 0775260

Smita Verma, 0768846

Brij Sheth, 0769631

St. Clair College of Applied Arts and Technology

Dab 304 Healthcare Analytics

Title

Admission analysis on mimic III

Group member:

Preetkumar SHAH, 0759151

Mitanshi Patel, 0775260

Smita Verma, 0768846

Brij Sheth, 0769631

**Table of Content**

[1) INTRODUCTION 3](#_Toc90039955)

[2) Related work 4](#_Toc90039956)

[3) Methods 4](#_Toc90039957)

[3.1) Data cleaning in Excel 4](#_Toc90039958)

[3.2) Data cleaning in Python 5](#_Toc90039959)

[3.3) Exploratory Data Analysis (EDA) 6](#_Toc90039960)

[3.4) Machine Learning 14](#_Toc90039961)

[3.5) Evaluating Matric and Result 14](#_Toc90039962)

[3.6) Machine Learning model Result 14](#_Toc90039963)

[3.7) Confusion Matrix 15](#_Toc90039964)

[3.8) Roc Curve 16](#_Toc90039965)

[4) Contribution 17](#_Toc90039966)

[5) Conclusion 18](#_Toc90039967)

[6) References 19](#_Toc90039968)

# INTRODUCTION

“A good Healthcare is basic fundamental right of every human. During the Covid pandemic we have seen the worst situation in whole world healthcare system, the entire infrastructure collapse in a moment. Now, in the current situation majority of country government decided to come together and try to improve the healthcare infrastructure .”

Therefore we decided to use our analytics skill to identify unrecognised problem as well as solve issues that are already present. we all decided to work on MIMIC III Admission dataset. We select this dataset as from our knowledge every hospital or healthcare clinic have general problem like managing staff members, doctors and managing hospital resources. Sometimes they have proper resources but they don’t know how to and when they have to use it. This admission file contains the majority of data that can be used to solve this kind of problem. The admission file contains the data of more than fifty thousand patients admitted in hospital. This large amount of data is good for visualization and machine learning model. We are also working on predicting readmission rate in 30 days using Machine Learning model using Natural Language Processing (NLP). Here readmission means if any patients is visiting hospital for their treatment for the first time, after completing their treatment they come again in hospital for the same treatment that is readmission. We make Machine Learning model to predict this readmission rate.

# 2) Related work

The main purpose of this section is to display and compare related work of our project with another or previously done project which is completed by another group or any individuals.

What we have done in this project is completely unique. We could not find any work which is done by another group. So, we could not make any comparison.

# 3) Methods

Our datasets contains tremendous amount of data. we work on 2 different dataset file. We have applied lot of procedure in this project like data cleaning in Excel and python on both dataset(ADMISSION and NOTESEVENT), We applied data pre-processing, tokenizer and countvectorizer, etc.

## 3.1) Data cleaning in Excel

We used Excel to clean some of the data. The ADMISSION file contains a lots of data like Diagnosis report of patient, time series data and numerical data.

1. We find the lots of empty values in DEATHTIME column because death time is only available when patient died and if patient alive then row is empty. So, for cleaning purpose we fill the null value with “Alive”.
2. DIAGNOSIS column has lots of error like “?” in front of the values. So, we remove the “?” symbol from the values using Filter, find and replace feature.

## 3.2) Data cleaning in Python

In this section we showcase data cleaning that is done in python. The below data cleaning part only belongs to ADMISSION file.

1. In ADMISSION dataset, we remove LANGUAGE and RELIGION column from Dataset as it was not very important.
2. In ADMISSION dataset, we decided to fill the empty values in MARITAL\_STATUS with Unknown Status.
3. EDREGTIME and EDROUTTIME are the time that patient was registered and discharge from the emergency department So, we fill decided to fill empty value with the non-Emergency.
4. In NOTESEVENT dataset, we take only columns that have only DISCHARGE\_SUMMARY because we use this dataset to predict the readmission rate.
5. We use word\_tokenize to clean the Text column in NOTESEVENTS. The Text columns contains the Discharge Summary of patients that is written by their respective doctors.

## 3.3) Exploratory Data Analysis (EDA)

**Highest Admission in Hospital**

Chart, sunburst chart

Description automatically generated

Description: - By looking at this pie chart we can say that highest number of patients were admitted in EMERGENCY location followed by URGENT and the least one is ELECTIVE.

**Insurance used by patient**

A screenshot of a computer

Description automatically generated

Description: - This bar graph shows that mostly patients were admitted in hospitals have MEDICARE and PRIVATE insurance. Only 611 patients do not have any insurance or maybe there insurance company refuse to pay.

**Top 5 disease patients’ diagnosis with**

Chart, bar chart

Description automatically generated

Description: - In this bar graph we can observe that majority of patient in the hospital are diagnosed with PNEUMONIA and PEPSIS. Which means hospital management should put more focus in this. They should hire more experience doctors and should have enough resources to deal with this.

**Patient admitted in hospital month wise**

**Chart, bar chart

Description automatically generated**

Description: - The above bar chart depicts that majority of people are admitted in August and July followed by March and the lease number of people admitted in the February and November.

**Patient death in Admitted location**

Chart, bar chart

Description automatically generated

Description: - From above chart it can be said that majority of people died in Emergency ward. The lease on died in PREMATURE and PHYS REFERRAL or NORMAL DELI section. 22 patient died those were admitted from other Hospital. There is another interesting fact that 1 patient died while transferring from other Healthcare institute.

**Average Length of stay of the patient by Admitted Location**

Description: - Average length of stay in Admission Location. So majority of people stay in HMO REFERRAL/SICK FOLLOWED BY CLINIC REFERRAL/PREMATURE location. So, hospital should increase staff in this location and reduce the staff in TRANSFER FROM SKILLED NUR location.

**Average Length of stay of the patient by Disease**

Description: - Mostly patients stay in the hospital for longer time period are suffering from THORACIC DISSECTION and R/O MI followed by LIVER FAILURE.

**Death rate of patients by Diagnose diseases**

Description: - Here in pie chart we display the deaths of patients by disease. 29% of patients died from PNEUMONIA and 23% died from PEPSIS. The most interesting fact that we found here, mostly people are diagnose with PNEUMONIA and PEPSIS and died too.

## 3.4) Machine Learning

We use Machine Learning model to predict the readmission rate of only those patients who were admitted in EMERGENCY admission type. We use Logistic regression model. The primary reason behind using logistic regression is, our predicting / output variable is in the form of binary (0 or 1) form. This model is best it comes to describe the relationship between one dependent variable and one or more ordinal and nominal or independent variable.

## 3.5) Evaluating Matric and Result

Evaluation Metric is used to calculate capability of a model. The metrics we used are below with the result

* F1 Score : 73%
* Precision : 62%
* Recall : 59%
* Accuracy : 74%

## 3.6) Machine Learning model Result

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | Training Accuracy : 69%  Testing Accuracy : 71:  Validation Accuracy: 72% |

## 3.7) Confusion Matrix

Chart, treemap chart

Description automatically generated

In confusion matrix we get [9894 + 497] true prediction.

## 3.8) Roc Curve

Chart, line chart

Description automatically generated

ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. We get 71% accuracy in ROC curve.

# 4) Contribution

|  |  |
| --- | --- |
| **Name** | **Contribution** |
| Preetkumar Shah | * Perform Data cleaning on both dataset (ADMISSION & NOTESEVENT). * Perform Machine Learning Model and also applied Hyper Parameters. * Exploratory data Analysis in python * Presentation |
| Mitanshi Patel | * Finding and selecting dataset * Research on each and every columns of dataset. * Making weekly report and paper submission report. * Perform Exploratory Data analysis on Excel |
| Smita Verma | * Perform Data cleaning on both dataset (ADMISSION & NOTESEVENT). * Perform Machine Learning Model and also applied Hyper Parameters. * Exploratory data Analysis in python * presentation |
| Brij Sheth | * Finding and selecting dataset. * Research on each and every columns of dataset. * Making weekly report and paper submission report.   Perform Exploratory Data analysis on Excel. |

# 5) Conclusion

In a nutshell we can say that more patients were diagnoses by PNEUMONIA and SEPSIS as well as died too. So, hospital / Healthcare organization should hire more experience doctors and more powerful/Higher dosage medicine.

Increase staff in HMO/REFRELL and SICK and Emergency department and Thoracic Dissection.

# 6) References

<https://www.youtube.com/watch?v=sszwe7KlsSw>

<https://github.com/sukilau/mimiciii-project>

<https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/>

<https://towardsdatascience.com/logistic-regression-model-tuning-with-scikit-learn-part-1-425142e01af5>

<https://mimic.mit.edu/docs/iii/tables/admissions/>

<https://mimic.mit.edu/docs/iii/tables/noteevents/>