Project Proposal

**Team: Pinnacle**

Meghanjali Chennupati, Muralidhar Reddy Reddem, Keerthi Rayapati, Anil Babu Davuluri, Pradeep Reddy Yenreddy, Varun Teja Kolluru

**Problem Statement**: This project aims to accurately predict the length of the stay for each patient so that the hospitals can optimize the resources and function better.

The globe is confronting several health concerns. These vary from outbreaks of vaccine-preventable illnesses like **covid19, non-communicable diseases(cancer, heart attack, diabetics), the Global influenza pandemic (SARS, flu, ...etc.), Weak primary health care … etc**.

one of the major health hazards in history is the pandemic situation of COVID-19, where there is a massive loss of life across the world. Due to the spike in the covid cases day by day there is a lot scarcity of beds and oxygen cylinders in hospitals. This situation makes us to drive in finding some effective solution to determine the length of stay (LOS) of the person in the hospital. So, that the hospital management can better plan to get adequate and provide better resources and facilities for the people.

In day-to-day life health problem increase like diabetes, cancer, heart attack ...etc. these are non-communicable diseases. Hospital pre-appointment and booking of beds are difficult due to a lack of resources. In this scenario, it’s very difficult for non-communicable disease patients to get beds and pre-appointments in hospitals.

**Goal:**

As Data science is all about getting insights from the data and predicting the future to solve situations effectively and efficiently. However, to solve the problem by using data science we need to have previous data. We have collected this data from Vidya antic survey data including during the covid situation, general hospitals, special wings in hospitals...etc. We can analyze the data on the people who got admitted to hospitals and predict the Length of the stay of each person.

**How our data analytics project initiative benefits the healthcare industry**

This type of problem statement belongs to Health care analytics. The pressure on healthcare institutions to enhance patient outcomes and provide better care is growing rapidly. Despite the difficulty of this scenario, it also provides enterprises with a chance to significantly increase the quality of service by utilizing more value and insights they provided. Healthcare analytics is the study of data using both quantitative and qualitative methods, and strategies for examining patterns and trends in the collected data while utilizing healthcare management. The average length of stay for a patient is one of many performance indicators. Knowing the length of stay (LOS) in advance helps hospitals tailor their care plans to cut down on LOS and the number of infections among patients, employees, and visitors.

Covid-19 is recently one of the most neglected areas to concentrate on and has come under scrutiny due to the pandemic: healthcare management. Patient duration of stay is one crucial statistic to monitor and forecast if one wishes to increase the effectiveness of healthcare management in a hospital, even if there are many use cases for data science in healthcare management.

At the time of admission, this metric assists hospitals in identifying patients who pose a high LOS risk (patients who will stay longer). Once identified, patients at high risk for LOS can have their treatment plans improved to reduce LOS and lessen the possibility of infection in staff or visitors. Additionally, prior awareness of LOS might help with planning logistics like room and bed allotment.

The goal is to correctly anticipate the length of stay for each patient on an individual basis so that hospitals may utilize this data to better allocate resources and operate.

**Description of the data**: The data set is collected from hospitals across India. It is divided into training and testing data sets. The data set consists of 455495 observations for which patient length of stay can be predicted from 17 attributes. Our target variable is “Stay” and it is a multi-class classifier. It contains the range of days the person will stay in the hospital. There are 11 different classes in the target variable ranging from 0 days to more than 100 days.

**Data set overview :**

|  |  |
| --- | --- |
| Attributes  1.Case\_id  2.Hospital\_code  3.Hospital\_type\_code  4.City\_code\_hospital  5.Hospital region code  6.Availaible extra rooms in Hospital  7.Department  8.Ward\_Type  9.Ward Facility Code  10.Bed Grade  11.Patient\_id  12.city\_code\_patient  13.Type of admission  14.Severity of illness  15.Visitors with the patient  16.Age  17.Admission deposit  18.Stay | Description  Case id registered in Hospital  Unique code for the Hospital  Unique code for the type of hospital  City code of the hospital  Region code of the hospital  Number of extra rooms available in Hospital  Department overlooking the case  Code for the ward type  Code for the ward facility  Condition of bed in the ward  Unique patient id  City code for the patient  Admission type registered by the hospital  Severity of illness recorded at the time of admission.  No of visitors with the patient  Age of the patient  Deposit at the time of admission  Patient length of the stay |

**Sources of the data**: The data is collected from Famous hospitals across all top cities in Punjab state in India. The is the patient’s historical data who got admitted to the hospital during covid. This is the data collected in the past two years.

Link to the data: <https://datahack.analyticsvidhya.com/contest/janatahack-healthcare-analytics-ii/True/#ProblemStatement>

**What we hope to Accomplish**: In this project in addition to the length of stay we want to further improvise by considering the age factor and severity of the disease of a person in determining the length of stay. So, we can effectively give solutions to the hospitals to increase their resources in hand if there is an increase of patients admitted into hospitals of a certain age. For example, in general, if the severity of the disease is extreme and if the patient’s age group is around 50 and more, there might be an increase in the chances of staying the person for more days in the hospital. So, if a large group of people ranging from age 50 or more is admitted into a hospital, then, hospital management can effectively manage the resources to tackle the situation.

**The logical plan that we are attempting to try:**

**Data Preprocessing Cleaning:**

**Step-1:** First we will do the basic exploration and cleaning of the data set by checking the null values, dropping the duplicates, and getting more insights about the data.

**Data Transformation** and **Feature Engineering**:

**Step 2:** We want to do feature engineering for effective performance for our model by selecting the best predictors that are useful to predict the stay. In the data set, we have 17 attributes, however, these are basically related to two types of levels.

1. Patient level (Type of Admission, Severity of illness, Visitors with patient, Age, Admission Deposit)
2. Hospital level(Ward type, Department)

Based on our data set, we thought the above attributes would help in better prediction. We will explore further by using feature engineering.

We will split the data into training and testing data. In our data set, there is no data imbalance. So we can continue by skipping the data imbalance addressing techniques.

**Data Modeling**:

**Step-3**: In the existing project they use Naive Bayer’s classifier, XG boost techniques.

However, we want to try some of the multi-class classification algorithms like

1. Decision trees

2.KNN

3. Random Forest

4. AdaBoost

5. Gradient Boosting

6.XGboost

7. Artificial Neural Network

**Data visualization**:

**Step-4**: The predictions of our models and how the past data is present, and factors are influenced by plotting the graphs for better understanding using the matplotlib, and seaborn libraries.

**Evaluation of model by metrics**:

In the existing project (naive bayer’s ), they predicted the length of stay for a given case id. They haven’t considered any further information like age, severity ...etc. regarding any predictors present in the data. So, this gave us the intuition to further improvise this by giving some solutions, and suggestions for health care.

In the navies Bayer’s classifier there few drawbacks in model

* Naive Bayes assumes that all predictors (or features) are independent, which happens to open in real life. This limits the applicability of this algorithm in real-world use cases.
* This algorithm faces the ‘zero-frequency problem’ where it assigns zero probability to a categorical variable whose category in the test data set wasn’t available in the training dataset. It would be best if you used a smoothing technique to overcome this issue.
* Its estimations can be wrong in some cases, so you shouldn’t take its probability outputs very seriously.

Based on metrics and cost evaluation of the previous model are not enough to address the problem. To increase metrics and cost apply a few models (KNN, Decision Tree, XGboost, ANN, AdaBoost). We can use metrics like Accuracy, Precision, Recall, F1-score, the ROC curve, and PR curve by exploring the threshold values to see which model is good at fitting our data set.

Intuition-based Artificial neural network is the best model based on the weights and hidden neurons they give the best metrics. This model is best helpful to government hospitals, private multiple specialty hospitals (multiple branches), small hospitals ...etc. They can analyze the estimation of new patient stay period, so they can increase or decrease the beds in hospitals.

**Extension to this project**: From the new patient stay period we can analyze the how much percentage of beds to increase in the hospital.