**Predicting the chances of survival of patients**

**Team members:**

* [Hymavathi Gummudala](https://iu.instructure.com/groups/824430/users/6679250)
* [Lakshmi Mounica Kalahasti](https://iu.instructure.com/groups/824430/users/6700966)
* [Sami King](https://iu.instructure.com/groups/824430/users/6109289)
* [Shashikant Reddy LNU](https://iu.instructure.com/groups/824430/users/6682725)
* [Mansi Pagare](https://iu.instructure.com/groups/824430/users/6698729)
* [Kanishkaa Sharma](https://iu.instructure.com/groups/824430/users/6689759)
* [Rohith Sureddi](https://iu.instructure.com/groups/824430/users/6693428)

**Introduction:**

In 2020 there was a total of 33,356,853 admissions to the hospitals in the United States.[[1]](#endnote-1) Understanding the overall condition of a patient’s health in a quick scan is crucial for health care workers. Several research over the years has been done to evaluate the attributes which are significant in specifying the chances of survival of a patient after getting admitted to a hospital. ICU’s mostly have insufficient medical data of incoming patients and the patients getting admitted to the ICU’s are majorly unable to provide them with their chronic condition information. We are trying to indicate the chances of survival of a patient based on their chronic health evaluation. The dataset includes the highest and lowest values of different tests and conditions within the first 24 hours of patient’s admission. It comprises of blood tests like sodium – potassium concentration in serum, blood pressure values, glucose estimate, heart rate count etc. Moreover, the dataset includes other attributes like body mass index, pH value of blood and genotypic sex of patient, all of which are notable characteristics in determining the chances of survival of a patient.

**Aim:**

The aim of this study is to identify and predict severity of illness of patient, chances of survival using patient’s demographic data, medical history and EHR records.

**Purpose:**

The purpose of our study is to determine the expansion and importance of the attributes that might indicate chances of survival for patients with illness. We plan to do so by taking the patient's data, medical history, and considering current conditions. Enable faster diagnosis, prediction of patient’s condition, health delivery protocols for chronic conditions by data analysis & interpretation.

**Methodology:**

1. Type of Analysis-
   1. Quantitative: Quasi-experimental research
   2. Classification: Naïve Bayes algorithm
   3. Decision tree Classification
   4. Imbalanced Classification
   5. Artificial neural network algorithm with Sigmoid activation function
2. Data Collection –

The dataset was found on Kaggle, a community-based repository of datasets.(<https://www.kaggle.com/datasets/sadiaanzum/patient-survival-prediction-dataset?select=Dataset.csv>). This dataset has a total of 186 columns and 91,714 rows. The rows indicate a patient, then the columns represent the patient’s characteristics and EHR data.

1. Storage and Extraction –

We first determined which attributes were viable for our prediction. The first step of this was to determine which attributes of the original 186 columns were pertinent to our prediction. We began this process by dividing the columns amongst the 7 team members. Each member was responsible for noting what the columns represented and the impact of the column on the project’s aim. Each member is also responsible for documenting the total number of NA and blank values within their section. Collectively, the group decided that if a column has NA or blank values that account for more than 35% of the total number of values, the column would be deleted. We will attach the document with team members' notes on the data description and their decision whether to keep the column or not in the Final Project Proposal. Some of the chosen attributes thus far are as follows:

1. Patient\_id
2. Hospital\_id
3. Hosptal\_death
4. Age
5. Ethnicity
6. Gender
7. Icu\_stay\_type
8. Aids
9. Cirrhsis
10. Diabetes\_melltius
11. Hepatic\_failure
12. Immunosuppression
13. Leukemia
14. Lymphoma
15. Solid\_tumor
16. Apache\_3j\_bodysystem (at the end of the dataset)
17. Apache\_2\_bodysystem (at the end of the dataset)

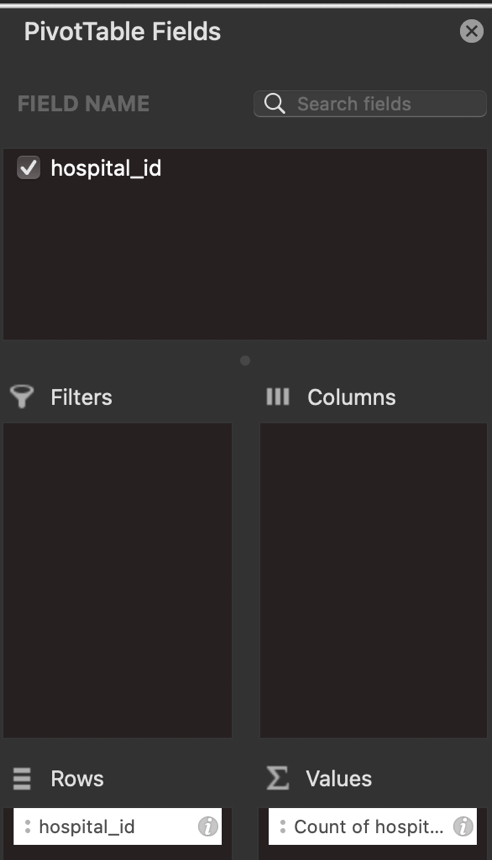
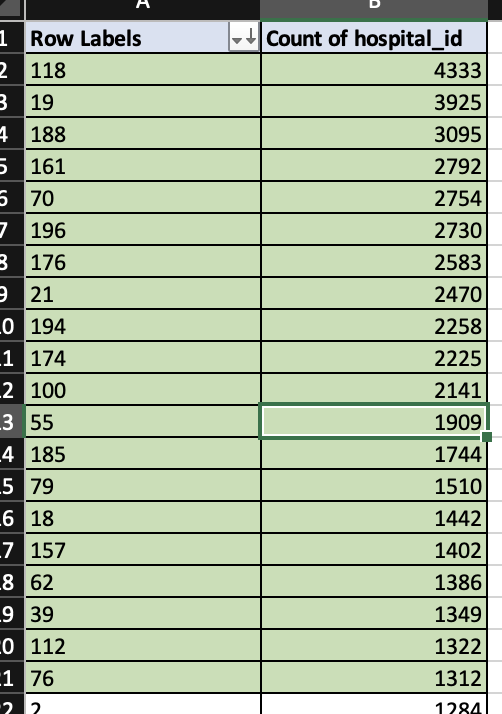
Storage: CSV file

1. Data Description / Data Cleaning –

After selecting the relevant attributes, we had a total of \_\_\_\_\_\_\_\_columns.

Next, we will condense the data set further. This is a preliminary example of our plan to reduce the number of hospitals represented in our data.

Sami will condense the dataset using the hospital\_id column. Sami will create a PivotTable like the point pictured below, in excel which will produce the count of hospital\_ids. This will allow us to condense our dataset and use hospitals with the highest number of entries. The initial reduction conducted below illustrates how the process will proceed after the data cleaning step. The preliminary cleaning reduced the number of rows to 44,683. This number might be different depending on the data cleaning process.



Row Labels = hospital\_id

1. Statistical Analysis -
   1. Model development and building
   2. We predict that we will be using Excel and Python in conjunction with Jupyter lab for our data analysis.
2. Tools –
   1. Tableau (visualizations)
   2. Excel
   3. JupyterLab

**Hypothesis:**

1. Null Hypothesis: According to our analysis and algorithms, independent features don’t have relations to the patient's health status and these features don’t predict the negative effect on patient health h status.
2. Alternate hypothesis: According to our analysis and algorithms, independent features have relations with the patient's health status and these features predict the negative effect on patient health status.

**Deliverables:**

1. To predict the effects that various factors will have on patients admitted to intensive care units
2. Using many variables, including the patient's medical history and current medical conditions, we will estimate the acceptance ratio and rejection rate.
3. Using a variety of techniques and statistical analysis, we will decide whether the result is acceptance or rejection.

**Results: -**

**1) Statistical analysis of features with Alternative Hypothesis represents the negative impact of patient health condition**

**2) Improving delivery and efficiency of medical care.**

**Team members Responsibility:**

|  |  |
| --- | --- |
| **Team Member Name** | **Responsibility** |
| Hymavathi Gummudala | * Data Cleaning & Feature selection – Rows 79-104 * EDA, Training the model , Model evaluation |
| Lakshmi Mounica Kalahasti | * Data Cleaning & Feature selection – Rows 157-186 * Exploratory data analysis |
| Sami King | * Data Collection * Data Cleaning & Feature selection – Rows 1-26 * Data Visualization |
| Shashikant Reddy | * Data Cleaning & Feature selection – Rows 27-52 * Model building, Model evaluation, and parameter fine-tuning |
| Mansi Pagare | * Data Cleaning & Feature selection – Rows 131-156 * Data Normalization |
| Kanishkaa Sharma | * Data Cleaning & Feature selection – Rows 105-130 * Data transformation & reduction |
| Rohith Sureddi | * Data Cleaning & Feature selection – Rows 53-78 * Model Building * R squared and f score model |

**Timeline:**

1. Dataset selection – 10/2/22
2. Dataset cleaning – 10/2/22 - 10/10/22
3. Hypothesis testing – 10/10/22 -10/15/22
4. Data analysis – 10/16/2022 - 10/22/22
5. Model building & visualizations – 10/23/22 - 10/29/22
6. Performance evaluation – 10/30/22 - 11/5/22
7. Project Demo – 11/6/22 - 11/14/22
8. Final project – 11/28/22 - 12/2/22

**References:-**

Kaggle. [www.kaggle.com](http://www.kaggle.com) ; <https://www.kaggle.com/datasets/sadiaanzum/patient-survival-prediction-dataset?resource=download>

1. Retrieved from: https://www.aha.org/statistics/fast-facts-us-hospitals [↑](#endnote-ref-1)