# Patient\_Survival\_Prediction

# Database

## Creating the Provisional Database

The original dataset had 91,713 rows of and 186 columns patient data. In order to prepare a sample dataset for testing our early machine learning algorithm (and begin exploring the data), we needed to perform a number of transformations. These included:

\*\*1. Dropping all of the columns related to lab work for each patient\*\* - although we will likely use some of these fields in our final model, we decided to eliminate these columns in order to reduce the size of our dataset to something some manageable for testing. Many of them had a high proportion of null values and would likely be dropped eventually. This step reduced the number of fields to 110.

\*\*2. Dropping all columns with more than 5000 missing values\*\* - because so many fields were missing large numbers of values, they needed to be removed. The threshold of 5000 was selected after trying cutoff values higher and lower than that amount. We felt that a threshold of 5000 still gave us a significant remaining dataset to work with. This step reduced the number of fields to 73.

\*\*3. Encoding items indicated as 'object' data types\*\* - to prepare our dataset for our model, we encoded the seven fields that incliuded non-numerical data.

\*\*4. Dropping patient rows with missing data\*\* - after removing the columns with a high proportion of missing data, there were still patients with missing data sporatically throughout their record (which is reasonable considering how many measurements and labs are taken in an ICU). We considered using a data imputation method to replace the missing values, but decided to remove any patient with any missing data from the dataset. After this step we still have a sizeable dataset of 72,643 rows.

## Creating the Final Database

The provisional database worked well with our selected classification model(s). Based on the resulting output from the model and additional research, we began putting together the next iteration of the database to hopefully improve on the model's performance. After ERT process was complete, our final database contains 73,262 rows and 69 columns. The steps taken included:

\*\*1. Keeping steps 2, 3, and 4 from our provisional database work\*\* - maintaining these steps still made sense for our final model.

\*\*2. Dropping patient height and weight measurements\*\* - because the dataset includes a calculated body mass index (bmi) score we decided to remove the two factors that contribute to that score.

\*\*3. Dropping calculated scores for APACHE II, III, and IV\*\* - the Acute Physiology and Chronic Health Evaluation (APACHE) tool is used in hospital intensive care settings to predict the recuperation and mortality of patients and to classify them. The tool has been modified over the decades since the creation of the original version in 1981 and APACHE II in 1985, with various additions resulting in APACHE III (1991) and APACHE IV (2006). Since our dataset contained all of the factors that contribute to a patient's APACHE score, as well as the score itself, we needed to drop one. It was decided to eliminate the scores themselves and retain the factors (various vitals and labs) to maintain a high number of inputs.

\*\*4. Dropping unique identifiers\*\* - finally, we made the decision to remove each patient's unique identifiers (patient ID and encounter ID) to hopefully improve the algorithm's performance. A search of resources online indicated that this was typically a best practice (although certainly not unilaterally agreed upon)

## Creating Categorical Datasets

In addition to our final, complete database we wanted to create smaller datasets built around categories of patient information. The accuracy of our primary model was encouraging and we were curious if we could achieve a similar performance when using smaller amounts of more homogeneous data. To explore this possibility, we created the following categorical datasets:

- Demographic: includes basic patient information such as age, gender, ethnicity, ICU type, etc.

- APACHE covariates: list of factors that contribute to the APACHE scores

- APACHE comorbidities: multiple illness diagnoses

- All vitals: all of the vitals taken on the patient such as heart rate, respiratory rate, blood pressure, etc.

- First HOUR vitals: all of the vitals taken on the patient in the first hour of their ICU stay

- First DAY vitals: all of the vitals taken on the patient in the first day of their ICU stay