COMS- W4995 AML PROJECT
PROJECT DELIVERABLE #1 - PROJECT PROPOSAL
GROUP 3

Atharva Kulkarni (ak5070), Divya Tadimeti (dt2760), Xiaoyi Zhou (xz3126), Yolanda Zhu (xz3013), Yu Yen Liu (cl4202)

1. Background

Predicting the outcomes for patients in the Intensive Care Unit (ICU) is significant as it will provide guidance to doctors for determining the prognosis. This is especially important for heart failure (HF) patients, as some of them have complex underlying diseases. Even with modern development in medicine, we don't fully understand what determines whether an ICU-admitted HF patient will survive their hospital stay. To address this issue, our project aims to create and test machine-learning based models that can predict in-hospital mortality for HF patients. A better understanding of these variables and predictions can help doctors make more informed decisions.

2. Dataset Introduction

The dataset contains 3 kinds of variables: Demographic characteristics, vital signs and laboratory variables. The target variable is Outcome (in-hospital mortality), defined as the vital status at the time of hospital discharge in survivors and non-survivors: 0 stands for Alive and 1 stands for Death. Among 48 features in the raw dataset, 10 of them are categorical variables, the left are numerical variables.
Dataset: https://www.kaggle.com/datasets/saurabhshahane/in-hospital-mortality-prediction

3. Problem Statement

Based on the given clinical parameters, stratify patients into risk categories (e.g., low, medium, high risk) for a particular health event.
Other potential questions:

- Predict the 'outcome' for a patient based on the available features. This can be framed as a binary classification problem if the 'outcome' column has two distinct values (e.g., 0 and 1 for negative and positive outcomes, respectively).
- Predict a patient's gender based on the given clinical measurements.
- Predict the likelihood of a negative outcome (e.g., a cardiovascular event) based on a range of patient health indicators.

4. Proposed ML Techniques

The methodology to predict in-hospital mortality for HF patients is broadly divided into four subparts given below:

- **Dataset Processing:**
  The dataset has 1176 samples and 51 columns. Dataset processing involves few things, namely assessing if the dataset is imbalanced, the amount of missing data and how to deal with it. Standardizing/ Normalization, if necessary. Exploratory Data Analysis, finding the correlation between various features. We need to perform the above mentioned tasks to make sure that our data is ready for the steps mentioned below to be able to make informed predictions.

- **Feature Selection:**
  Within our dataset, the Recursive Feature Elimination (RFE) can be helpful by constructing numerous models and ranking the features, allowing us to zero in on those with the highest impact. Furthermore, given the mixed nature of our data, tree-based models could be useful. The inherent feature importance ranking provided by models like Decision Trees and Random Forests has given us insights into which variables play a crucial role in our dataset's predictive power. Lastly, considering the variety of continuous and categorical variables in our dataset, statistical tests like ANOVA and the Chi-squared test can allow us to ascertain the statistical significance of each feature in relation to our target. These techniques ensure that our model is not just theoretically sound, but is also tailored to the unique characteristics and intricacies of our dataset.

- **Model Selection:**
  For predicting in-hospital mortality of patients and stratifying them into risk categories, Logistic Regression is a straightforward choice for an initial binary classification task, classifying patients as survivors or non-survivors. It's interpretable and can be valuable for initial predictions. Decision Trees and Random Forests can be a next step as they capture complex relationships, making them suitable for both binary classification and risk stratification tasks. Lastly, we also plan to use Gradient Boosting models. We choose these models because they can handle both numerical and categorical data.

- **Model Evaluation:**
  We can evaluate the model performance using three methods namely Performance Metrics - a combination of metrics like precision, recall, F1-score, and the AUROC. Secondly, to ensure that our model doesn't just perform well on a specific subset of data, we will use k-fold cross-validation. And Feature Importance: Understanding which features play a pivotal role in predictions is vital. This will help interpret the model and identify critical clinical parameters for in-hospital mortality.