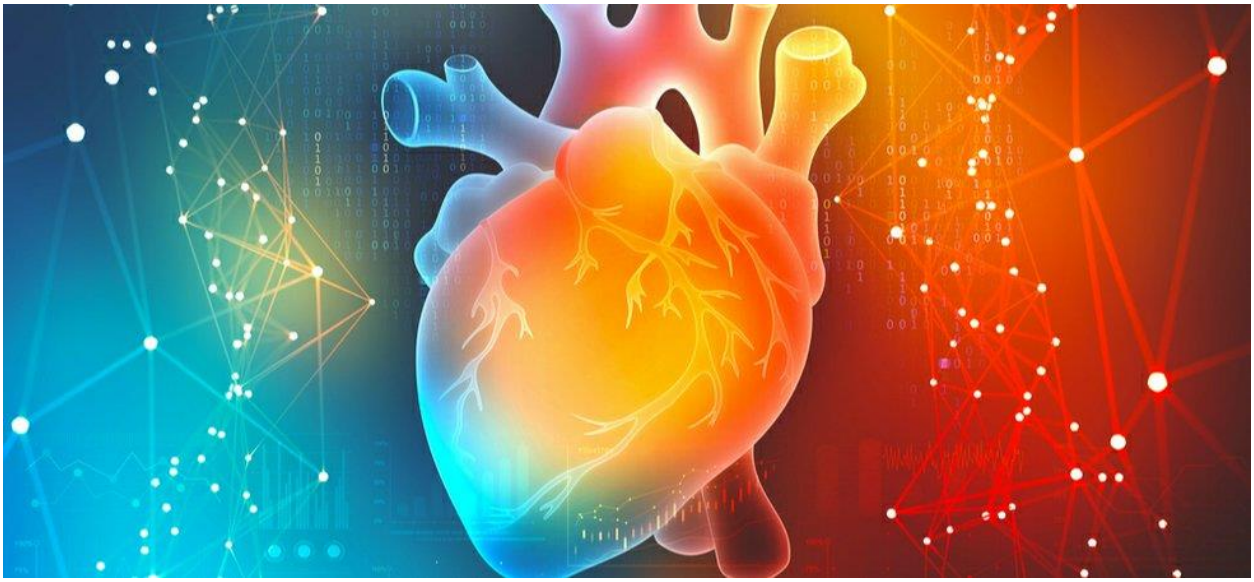
CARDIOVASCULAR PATIENTS HOSPITAL ADMISSION USING STATISTICAL & ML TECHNIQUES



DATA 607 – Statistical and Machine Learning

Project Proposal

Team

Carlos Magno Nascimento

Ekpo Archibong

Md Athar Khan

Md. Sajid

Shaila Hossain

March 25, 2023

University of Calgary

Dataset(s)

We have utilized the "Hospital Admissions Data" dataset (File size: 2.6 MB, Rows: 15757 K, Columns: 56) available in the Kaggle portal (https://www.kaggle.com/datasets/ashishsahani/hospital-admissions-data/discussion/302894?resource=download&select=HDHI+Admission+data.csv) in CSV format.

This data was collected from patients admitted over a period of two years (1 April 2017 to 31 March 2019) at Hero DMC Heart Institute, Unit of Dayanand Medical College and Hospital, Ludhiana, Punjab, India. Specifically, data were related to patient's date of admission; date of discharge; demographics, type of admission; patient history, and lab parameters such as hemoglobin (HB), total lymphocyte count (TLC), platelets, etc. Other comorbidities and features (28 features), including heart failure, STEMI, and pulmonary embolism, were recorded, and analyzed. The outcomes indicating whether the patient was discharged or expired in the hospital were also recorded.

table_headings.csv - This data table has the descriptive headlines for all columns for the HDHI Admission data file.

Research Questions

- Which independent variables are the most important to predict heart failure and AKI?
- Which statistical method best predicts heart failure, AKI, Duration of stay and Outcome in terms of misclassification rate/MSE?
- In terms of misclassification rate which model best predicts the 3 levels of the variable OUTCOME (Expiry, Discharge, DAMA)?

Data Wrangling & Analysis Procedure

For data wrangling and analysis/visualization we will be conducting below executions using Python:

- Data type conversions
- Sorting all the datasets for systematic analysis.
    - Deleting non-required columns as there are many categorical variables, null values.
    - Format checking
- Cross-check for duplicates.

We are planning to utilize the following techniques:

- Feature Scaling and KNN
- Decision Trees, Random Forest, Boosting, BART
- Principal Component Analysis
- Regression, QDA, cross validation
- Grid SearchCV and RandomizedSearchCV

Further techniques/questions will be included/removed in the final analysis and an accompanying report will be provided at the end of the project work.