**Survival analysis of heart failure patients in Faisalabad Institute of Cardiology**

# Sindhu Bhattarai, Sean Torres, Dingyi Duan

Shiley-Marcos School of Engineering, University of San Diego

# Abstract

Everything year approximately 17 million people are killed by cardiovascular diseases. Heart failure is one of the most exhibits and can be predicted using electronic medical records that contain different symptoms and laboratory test values. We observed 299 patients collectively who were admitted to the Institute of Cardiology and Allied hospital Faisalabad-Pakistan during April-December (2015) and applied data mining skills with several machine learning classifiers to predict patients’ survival and find the most determinant factors of said target.

The C5.0 and CART models produced accuracy scores of approximately 86%; whereas logistic regression and random forest models performed approximately [add]. Implementation of Naïve Bayes classification in conjunction with the neural network model yielded [add], respectively. We also present the model evaluation tables to cross-validate and compare each model to further explore the limitations and advantages.

*Keywords:* heart failure, machine learning, data mining, model evaluation

Table of Contents

[add last after all the other necessary contents are complete]

# Methodology

The original data set was obtained from UCI Machine Learning Repository (Ahmad, et al., 2019) and imported, read, and analyzed in Jupyter Notebook using Python. After confirming that our data set has 299 records and 13 attributes, we started exploratory data analysis (EDA) and performed preprocessing. Lucky for us, there was no null or missing data detected in the data set. So, after we use the first and third interquartile to remove the outliers, we went on with data processing. Among the 105 women and 194 men, who were all over the age of 40, are evaluated with anemia (decrease of red blood cells or hemoglobin, Boolean), high blood pressure (if the patient has hypertension, Boolean), creatinine phosphokinase (level of the CPK enzyme in the blood in mcg/L, Continuous), diabetes (if the patient has diabetes, Boolean), ejection fraction (percentage of blood leaving the heart at each contraction, continuous), platelets (platelets in the blood in kilo platelets/mL, continuous), serum creatinine (level of serum creatinine in the blood in mg/dL, continuous), serum sodium (level of serum sodium in the blood in mEq/L, continuous), smoking (if the patient smokes or not, Boolean), time (follow-up period, discreet), and our target DEATH\_EVENT indicates whether the patient was deceased or not. For dimensionality reduction, we first eliminated the “time” variable due to its increments regardless of DEATH\_EVENT. Then we use multicollinearity to find which attributes are more correlated to each other and take them out of our predictor variables. [Add standardization/normalization step after proven to be necessary for other models]. The data set was further subjected to a 75:25 train-test split ratio to build the models and use various model evaluation metrics and cross-validate to evaluate the models’ performance.