

**Design Document**

**Introduction**

**Scope**

The goal of my study is to assess the likelihood of a death by heart failure event, based on specific medical history. The conclusions of this study can be used to help the medical staff in assessing the severity of patients with cardiovascular diseases.

**Overview**

For this study, I will be using a heart failure clinical records dataset, which contains a medical history and details about patients who have experienced this medical condition to reach the goal I set above.

**Dataset review**

Row:

* Number of patients – There are 299 patients examined in the study.

Columns:

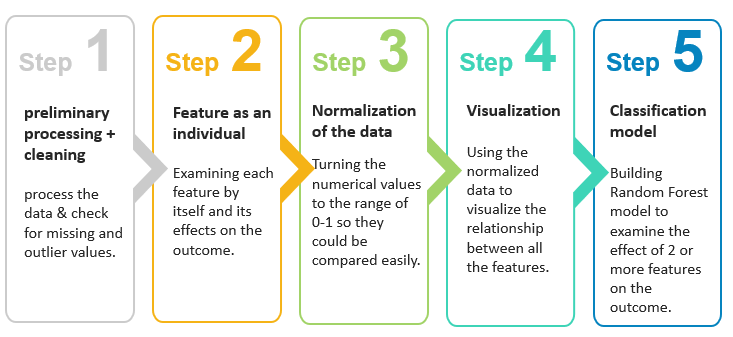
* Age – The patient's age in years.
  + Data type – Continuous.
  + Number of values – 299.
  + Null values – None.
* Anemia – Whether the patient has anemia or not.
  + Disorder in which the amount of red blood cells or the amount of hemoglobin in the blood has decreased to below normal levels.
  + Data type – Binary.
  + Number of values – 299.
  + Null values – None.
* Creatinine Phosphokinase – The enzyme levels in the patient’s blood.
  + This enzyme can indicate damage to the heart muscles once its values are above the normal range (for adults the range is 43-150).
  + Data type – Continuous.
  + Number of values – 299.
  + Null values – None.
* Diabetes – Whether the patient has anemia or not.
  + Disease in which the blood glucose levels in the body are too high.
  + Data type – Binary.
  + Number of values – 299.
  + Null values – None.
* Ejection Fraction – The Percentage of the blood that the left ventricle pumps out with each contraction.
  + Proper levels of EF are 53 (52 for women) to 73 percent.
  + Data type – Continuous.
  + Number of values – 299.
  + Null values – None.
* High blood pressure - Whether the patient has tendency to having high blood pressure.
  + Data type – Binary.
  + Number of values – 299.
  + Null values – None.
* Platelets – The number of platelets per microliter in the patient blood.
  + The normal number of platelets in the blood is 150,000 to 400,000 platelets per microliter.
  + Data type – Continuous.
  + Number of values – 299.
  + Null values – None.
* Serum creatinine – The creatinine levels in the patient’s blood.
  + A creatinine level of greater than 1.2 for women and greater than 1.4 for men may be an early sign that the kidneys are not working properly.
  + Data type – Continuous.
  + Number of values – 299.
  + Null values – None.
* Seron sodium – The sodium levels per liter in the patient’s blood.
  + A normal blood sodium level is between 135 and 145 milliequivalents per liter.
  + Data type – Continuous.
  + Number of values – 299.
  + Null values – None.
* Sex – Whether the patient is male or female.
  + Data type – Binary.
  + Number of values – 299.
  + Null values – None.
* Smoking – Whether the patient is a smoker or not.
  + Data type – Binary.
  + Number of values – 299.
  + Null values – None.
* Time – The number of minutes that gave passed between commencement and termination (either healing or death) of treatment of the patient.
  + Data type – Continuous.
  + Number of values – 299.
  + Null values – None.
* Death event – Whether the patient died at the end of the treatment or not.
  + Data type – Binary.
  + Number of values – 299.
  + Null values – None.

**Working Prosses & Description**

**Language & packages**

1. Language – Python.
2. Packages that were used in this project:
   1. Information management – Numpy, Pandas, Scipy.
   2. Visualization – Matplotlib, Seaborn.
   3. Classification models – Random Forest, K fold validation.

**Working prosses**



**Classification model – Random Forest**

1. In this project I used a decision tree classification model for further understanding of the impact of the features combined.
2. In order to determine the number of trees in the model (n\_estimators), I used k fold validation in the range of 1-150 trees.
3. The most efficient number of trees (variable ne\_max) was used in building the model.
4. The dataset was split to train and test groups in ratio of 0.9:0.1 respectively.
5. I used the attribute “stratify” to make sure that the ratio of the “DEATH\_EVENT” values (0 or 1) in the train group will be equal to the ratio in the test group.
6. The model score using the test group showed 98% accuracy.

**Conclusions**

**Analyzing the data**

1. Time has a direct impact on outcome – as time passes by, the risk of death from heart failure is getting higher.
2. Ejection Fraction – when EF value is lower than 30, the risk of death is higher. That being said, the EF value among about half of the death cases is higher than 30, so that conclusion isn’t two way, meaning a patient in risk of death can have EF value higher than 30.
3. Serum Creatinine – when SC value is higher than 1.5, the risk of death is higher. That being said, the SC value among some of the death cases is lower than 1.5, so that conclusion isn’t two way, meaning a patient in risk of death can have SC value lower than 1.5.
4. Age – there is a higher risk of death among the 80+ age group. However, the sample of that age group is very small, while most of the patients in the dataset are in the 50-70 age group. Therefore, while there are more death cases among patients in their 80s and 90s based on the dataset, this conclusion is inconclusive due to the lack of samples in this age group.
5. Ejection Fraction & Serum Creatinine – the random forest results showed that when the values of the EF are above 30 (which based on section 2, are values that both death and recovery cases share) and the SC values are above 1.75 (which based on section 3, can put a patient in a higher risk of death), there is a higher risk of death from heart failure. This conclusion can indicate on the importance of the SC feature and its effect on the outcome as shown at section 3.
6. 40-70 age group – as shown by the random forest results, most of the young patients (ages 40-70s) who died after a heart failure, follow the same criteria; Ejection Fraction values between 27.5 and 32.5, Serum Creatinine values above 0.85, and Serum Sodium above 135. If a patient meets all of the above criteria, he/she is in higher risk of death.

**Further research**

1. The conclusions that presented in this project are based on the dataset.
2. As mentioned, the sample group in the dataset is small (about 300 patients tested). Therefore, it is difficult to draw conclusions regarding the sample population.
3. That being said, the conclusions in this project can be used for further research.
4. Additional data that can be collected as a part of a follow up research is healthy patients’ medical record. Since in this dataset a lot of the values were overlapping among recovery and death cases (since all of the patients had heart event), examining medical record of healthy patients will provide deeper understanding on the features’ effect.