**Project Title: Predicting the Risk parameters for Cardiovascular patients**

**Introduction**

Cardiovascular Disease often used interchangeably with “heart disease”, generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves, or rhythm, also are considered forms of heart disease.

The purpose of this project is to predict the effects of different parameters recorded in the data to predict mortality of the patient. By predicting so the physicians can determine high risk patients and can take better care of them thus helping them survive.

**Statistical / Hypothetical Question**

Following hypothesis was tested.

* Ejection Fraction is lower for Dead Patients vs Alive Patients. Hence it is an indicator of high risk for Mortality
* Serum Creatinine is higher for Dead Patients vs Alive Patients. Hence it is an indicator of high risk for Mortality

**Outcome of your EDA**

We do see some relation between some parameters/variables on the mortality.

The most important are

1. Ejection Fraction - If the % age is lesser, Risk increases.

2. Creatinine Level - Wit increase in level, Risk increases.

3. Sodium Level - With decrease in level, Risk increases.

4. Age - Older the patient, higher the risk

The other factors which I observed, did not show much effect.

I was able to support our hypothesis with the tests.

I also was able to generate prediction model (a logistic regression model), and found it be 76.25% accurate with the current dataset.

**What do you feel was missed during the analysis?**

Another factor which I did not plan earlier was

1. Time in days between follow up visits.

While analyzing separately, I did notice time (in days) between follow up visits was making the risk lower. It is anyways evident that if you go on regular follow ups, you will be able to know the problem before its too late and you may have the opportunity to act to decrease the risk.

**Were there any variables you felt could have helped in the analysis?**

Some of the variables, which I felt should have helped in analysis was the region, country and some more variables on eating habits, cholesterol levels, family history etc. which would have made the study more accurate. Using them would have taken care of region specific as well as habit specific filtering and accurate predictions.

**Were there any assumptions made you felt were incorrect?**

There were some concerns regarding the sample size of the data. Since the data was from a single location, there might be some other factors in play which can be due to habits of people to that specific region or part of the world and may not be reflecting in the data. If we get more geographically separated data, we might come up with better model. This model may be more effective in the region where data comes from but our assumption that it will work for all might be wrong.

**What challenges did you face, what did you not fully understand?**

I had some challenges around the Hypothesis Testing and choice of the test statistic. I chose the difference between the means of the two groups as my main test statistics, however I still think could I use some other comparisons too as test statistics such as standard deviation or chi squared based tests.

I feel still as a novice and learning as I read and practice with different datasets.

**References:**

<https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118#:~:text=The%20term%20%22heart%20disease%22%20is,pain%20(angina)%20or%20stroke>.

<https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>

<https://www.healthline.com/health/creatinine-blood#results>

Authors of this Dataset

Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020).

Discover Statistics Using R, Andy Field | Jeremy Miles | Zoe Field

R for Everyone, Jared P Lander

Think Stats, Allen B Downey