**Name** : Abhigyan Misra  
**Date** : 11/15/20  
**Title** : Predicting the Risk parameters for Cardiovascular patients

**Section 1**

* **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.

* **Research questions**

The attributes or parameters which I would be looking at specifically for analysis on this project are

* Age of Patient
* CPK Levels
* Ejection Fraction
* Platelets
* Serum Creatinine
* Serum Sodium

Based on these parameters, I would like to see which one of them effect the mortality the most. If we are able to understand the thresholds and come up with a model for such parameters which can predict if the patient is high risk or not, it would help to take care of them at right time.

I will be using death event as an outcome variable, which may be considered as High-Risk parameter too, to define the model.

* **Approach**

To start with, I will first try to segregate the patients based on death event and plot the histograms for the 6 parameters that I will be analyzing.

I would also like to see if we have some strong correlation of these parameters with the death event using the scatter plots.

We may also try to see if the data is clustered or not and based on the same can decide what type of model would be best for this purpose.

The dataset would be split as train and test data with split ratio of 80% and then we may try to check the accuracy of the model as well.

* **How your approach addresses (fully or partially) the problem.**

The approach, I would say only partially address the problem due to limited data which is region specific and not for the whole world. There may be some influence due to specific local region due to eating habits, stress levels, or some other unknown parameters which are not captured in this dataset.

However, if we get more data from different regions of the world, we may be able to come to a better way of predicting the High-Risk patient. The accuracy that we may calculate will only be based on this limited data.

This project may act as steppingstone for further analysis with more similar datasets available across different hospitals from around the world.

* **Data**

The dataset contains the medical records of 299 heart failure patients collected at a Cardiology Institute and at an Allied Hospital in Asia region, during April–December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. All 299 patients had left ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) classification of the stages of heart failure

The dataset contains 13 features, which report clinical, body, and lifestyle information (See Table [below](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5#Tab1)), that we briefly describe here. Some features are binary: anemia, high blood pressure, diabetes, sex, and smoking. The hospital physician considered a patient having anemia if hematocrit levels were lower than 36%. Unfortunately, the original dataset manuscript provides no definition of high blood pressure, however assuming standard measurements of 120/70 as normal.

The below table lists the attributes in the dataset along with description, measurement scale and range of values that we might see in the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute Name | Description of Attribute | Scale/Measurement | Range Of Values |
| Age | Age of the patient | Years | [40,..., 95] |
| Anaemia | Decrease of red blood cells or hemoglobin | Boolean | 0, 1 |
| High blood pressure | If a patient has hypertension | Boolean | 0, 1 |
| Creatinine phosphokinase-(CPK) | Level of the CPK enzyme in the blood | mcg/L | [23,..., 7861] |
| Diabetes | If the patient has diabetes | Boolean | 0, 1 |
| Ejection fraction | Percentage of blood leaving  the heart at each contraction | Percentage | [14,..., 80] |
| Sex | Woman or man | Binary | 0, 1 |
| Platelets | Platelets in the blood | kiloplatelets/mL | [25.01,..., 850.00] |
| Serum creatinine | Level of creatinine in the blood | mg/dL | [0.50,..., 9.40] |
| Serum sodium | Level of sodium in the blood | mEq/L | [114,..., 148] |
| Smoking | If the patient smokes | Boolean | 0, 1 |
| Time | Follow-up period | Days | [4,...,285] |
| (target) death event | If the patient died during the follow-up period | Boolean | 0, 1 |

* **Required Packages**

In this project, I will be utilizing several libraries and packages, some provided with R and some additional for data cleansing, plotting the data and visualizations, comparing the models etc.

Based on what plots and models, I may use below list of packages and more.

* Ggplot2
* Tidyverse
* Knitr
* caTools
* class
* psych
* **Plots and Table Needs**

Below is the list of plots and tables that I think will be needed to do the analysis and come up with a model

* Histogram
* Scatter plots for correlation
* Prediction Plots
* Accuracy Tables
* **Questions for future steps.**

For now, all my questions for future are based on how we can get more data to prove the study. What can be different data sources providing this information.

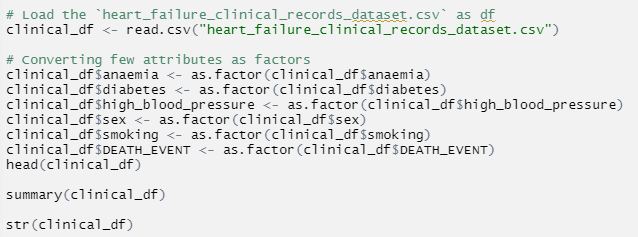
Is there a standard organization, could be WHO (World Health Organization), which collects all this kind of data and stores in a common repository or database?

Further, how valid is the data from non-developed or developing nations to be specific. This question comes to my mind because I do come from such a background and not sure about the validity of this data in a lot of scenarios. A lot of cases go unreported thus we miss the bigger chunk of data. Our whole study sits on this data and if this is error prone, then we may be analyzing wrong. There may be bigger efforts required to have the data collected properly at the source than generating or analyzing the model to predict more accurately.

**Section 2**

* **How to import and clean my data**

I have imported my data off Kaggle to my local computer and have imported the csv file. The csv file is kept in the same location as my RMD file.



I have also converted few attributes as factor as shown above. Those include – anemia, diabetes, high\_blood\_pressure, sex, smoking, DEATH\_EVENT.

I know that I must make sure that there are no holes in the data, because errors would be produced, and data would not be constructed correctly if there are holes. However the data seems pretty clean and only way I could check the data quality is by checking the possible values and their counts to match up the numbers and do a check if the columns have valid values.

Checked for NA values using is.na and did not find any.

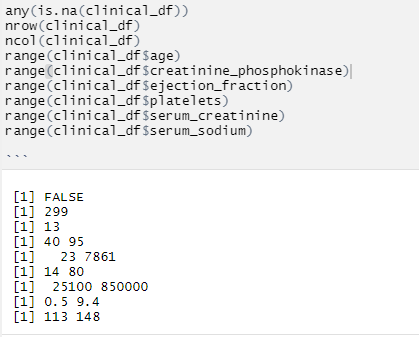
any(is.na(clinical\_df))

Checked the count of Records just to be sure. Was mentioned in the dataset as 299 patients which matched.

nrow(clinical\_df) - 299

ncol(clinical\_df) - 13

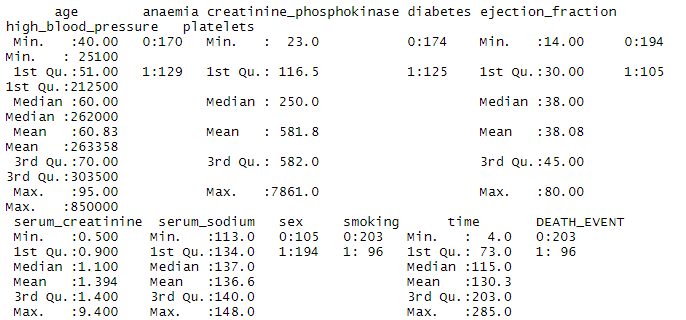
Also checked the range of values for numerical parameters. There are no negative or absurd values. Some are over the normal ranges for a fit person, but these variations may be legitimate due to person’s health.

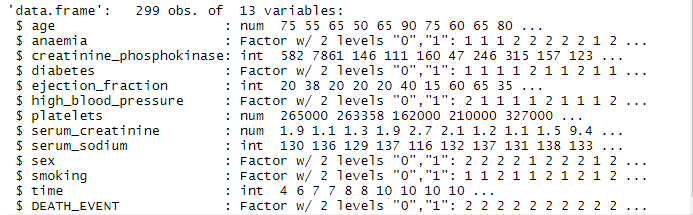


Overall, the Dataset was clean and Kaggle had a data rating of 10/10.

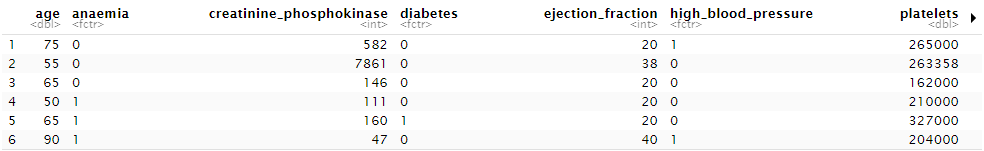
* **What does the final data set look like?**

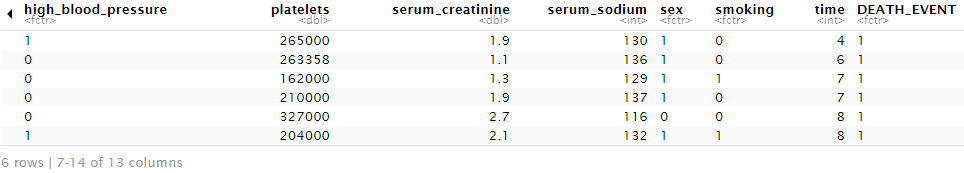
Here is the Summary & Structure of the data, which tells about how big the data is and what are the values.





Also printed the head for the data frame and below is the screenshot of the same.





* **What information is not self-evident?**

I hope to look at all the datasets and find common factors and relevant data that can be combined to answer questions not previously asked in the project. I also plan on using several different plots/graphs to help me visualize the data that will help with more technical parts. I may also consider creating a few different variables to getting a more detailed look at the data.

* **What are different ways you could look at this data?**

I might try to look at this data by controlling some of the parameters such as smoking, sex, diabetes, anemia which might show something specific to these. Though these parameters are not shown as significant but maybe they give some information which we are not able to see directly.

* **How do you plan to slice and dice the data?**

I plan to split the data using DEATH\_EVENT in case to see the effect of these parameters specifically on the mortality of the patient.

I would also be splitting the data into train and test to train and test my generated models.

I might further slice the data as needed for analysis based on the findings from previous steps.

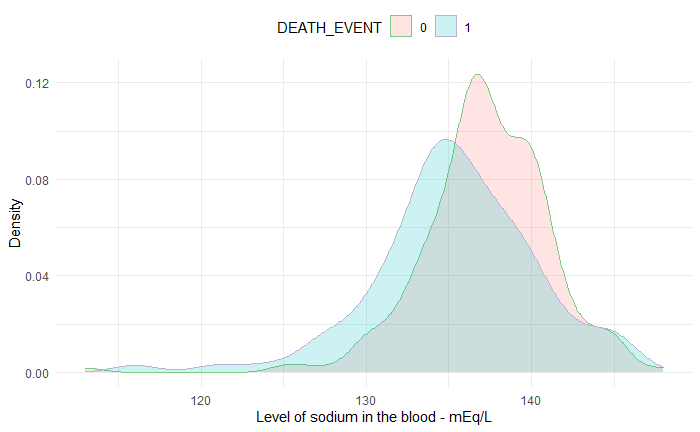
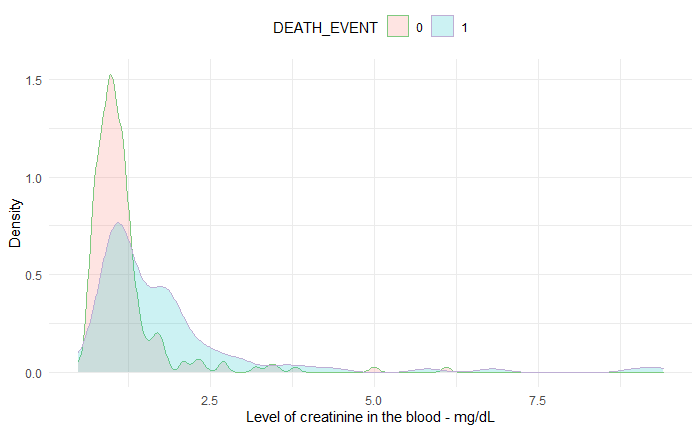
* **How could you summarize your data to answer key questions?**

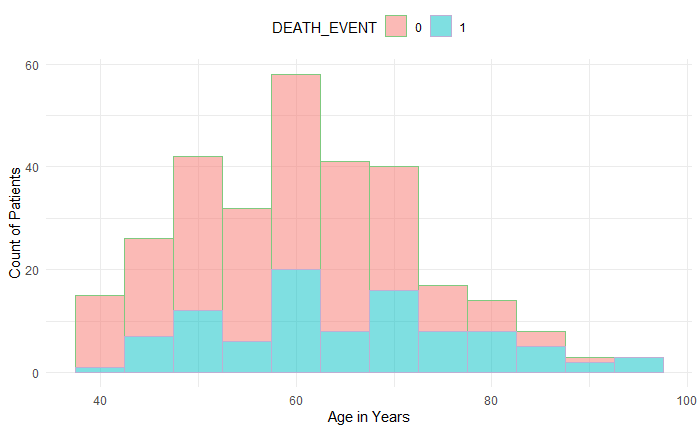
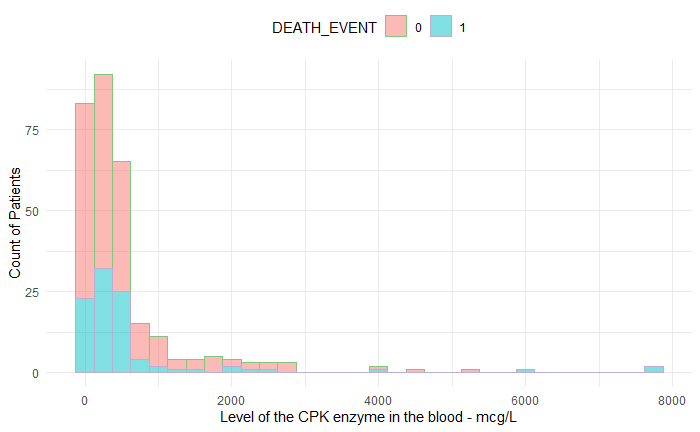
Based on my top question, the data would be able to tell me which parameters contribute more so to mortality after running a regression. So those parameters should be marked as high-risk parameters and should be closely watched for any fluctuations. Further I think that it would also be simple to calculate the percentage risk of mortality if I try feeding the values to the prediction model generated out of the data.

* **What types of plots and tables will help you to illustrate the findings to your questions?**

I would be using the Histograms and Density plots for patients who survived vs who did not survive to see the effect of such parameters on the mortality. Seeing the plots might make things clearer and some of the analysis could be made from the plot itself by observing which one’s show more effect on one’s mortality.

Putting some of the plots below.

* **Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.**

Yes, based on the significant parameters that would be found in the study, would like to generate a model, which could predict the risk of mortality.

* **Questions for future steps.**

Since being a first such kind of project for me, I am still not sure whether the prediction which I would make be right or not. To validate the same, I would need to run this prediction model on various other datasets and may be needed to train this again once I get a large dataset with similar information.

**Section 3**

**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.

**The problem statement you addressed.**

Below are the key questions that I looked for during my analysis on the data.

1. What parameters effect the Mortality rate and how?
2. Which parameters are more crucial in evaluating the risk for the patient?
3. Can we get a Model to predict the mortality?
4. What is the accuracy of the model based on the data available?

**How you addressed this problem statement**

To address the problem, I first plotted the Histogram plots and the Density plots of the mentioned attributes (Ejection Fraction, Age, CPK Levels, Creatinine Levels, Sodium Levels, Platelets Counts) with the Death Event. I was able to identify the attributes which affected the mortality the most.

Later I split the dataset into train and test datasets and tried to generate models for the death event as outcome and combination of other attributes as input variables. I came up with multiple models and then compared their accuracy and tried fitting the model to understand them.

I was able to come up with a logistics regression model with the most accuracy.

**Analysis**

The analysis clearly showed how the Ejection fraction, Age, Creatinine Levels and Sodium Levels affected the mortality rate of a patient. Age was a variable which we could understand without analyzing, however it clearly showed the same on the plots.

Ejection Fraction and Creatinine Levels came out to be more fragile or riskier variables, which affect the mortality rate the most.

Sodium Levels also had some effect on the mortality but not same as Ejection Fraction and Creatinine Levels.

I first tried generating the model for the death event using these 4 variables however, while looking at a test model with all variables from the data frame, time(in days) between follow up visits came up as significant variable and then I generated another model using the same.

While checking the accuracy and fitness of the model, this model with time as input was better than the previous model.

**Implications**

One of the implications that came out of the analysis was that, though we were not observing the time between follow ups, but it showed affect on the mortality rate directly. One may understand without even looking at the data that if a patient does regular and frequent follow up visits then he surely can control his vitals thus having a positive effect on his survivability. This was also shown when we tried generating the models and time between follow ups came up as significant parameter.

Another such parameter was Age, which one may anyways understand without even analysis that as the age increases the risk of mortality increases. So, although we plotted and saw it, it directly implies that as age will increase there is always increased risk of mortality.

**Limitations**

The study or analysis had some limitations due to the size of data being analyzed. The dataset I chose only had 299 data points, it would have been better, if a larger number of patient data were used. This can be overcome by going through some other health agencies which publish such data to have a better model.

Further the data collected was from a single location or region thus there could be some effects due to region specific parameters such as eating habits, smoking habits, daily schedules etc.

**Concluding Remarks**

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

Another factor which I did not plan earlier was

1. Time in days between follow up visits.

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

We were able to generate multiple prediction models and we found one the models (a logistic regression model), which is shown as "lrmodel.3" in the analysis, to be more accurate and fit than the others.

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.

**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>

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