**Introduction:**

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worlwide. Heart failure is a common event caused by CVDs, which could be derivative of 12 different features that may be used to predict mortality by heart failure.

Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

In this context, electronic health records (EHRs, also called medical records) can be considered a useful resource of information to unveil hidden and non-obvious correlations and relationships between patients’ data, not only for research but also for clinical practice and for debunking traditional myths on risk factors. To this aim, several screening studies have been conducted in the last years, covering different conditions and demographics and with different data sources, to deepen the knowledge on the risk factors. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

Machine learning applied to medical records, in particular, can be an effective tool both to predict the survival of each patient having heart failure symptoms and to detect the most important clinical features (or risk factors) that may lead to heart failure.

**DataSet:**

I analyzed a dataset containing the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), 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, that is briefly described here. Some features are binary: anaemia, high blood pressure, diabetes, sex, and smoking.

Anaemia refers to a condition when if haematocrit levels were lower than 36%.

The creatinine phosphokinase (CPK) states the level of the CPK enzyme in blood. When a muscle tissue gets damaged, CPK flows into the blood. Therefore, high levels of CPK in the blood of a patient might indicate a heart failure or injury.

The ejection fraction states the percentage of how much blood the left ventricle pumps out with each contraction.

The serum creatinine is a waste product generated by creatine, when a muscle breaks down. Especially, doctors focus on serum creatinine in blood to check kidney function. If a patient has high levels of serum creatinine, it may indicate renal dysfunction.

Sodium is a mineral that serves for the correct functioning of muscles and nerves. The serum sodium test is a routine blood exam that indicates if a patient has normal levels of sodium in the blood. An abnormally low level of sodium in the blood might be caused by heart failure.

The death event feature, that we use as the target in our binary classification study, states if the patient died or survived before the end of the follow-up period, that was 130 days on average. Regarding the dataset imbalance, the survived patients (death event = 0) are 203, while the dead patients (death event = 1) are 96.

The annotations of dataset are here as follows:-

1) Age: Age of the patient

2) Anaemia: Decrease of red blood cells or hemoglobin (boolean)

3) CPK: Creatine phosphate kinase level in the blood (mcg/L)

4) diabetes: If the patient has diabetes (boolean)

5) ejection\_fraction: Percentage of blood leaving the heart at each contraction (%)

6) high\_blood\_pressure: If the patient has hypertension (boolean)

7) Platelets: Platelets in the blood (kiloplatelets/mL)

8) Creatinine: Level of serum creatinine in the blood (mg/dL)

9) Sodium: Level of serum sodium in the blood (mEq/L)

10) Sex: Woman or man (binary)

11) Time: Follow up period (days)

12) Event: Whether the patient survived or not (binary)

13) Smoking: If the person smokes (binary)

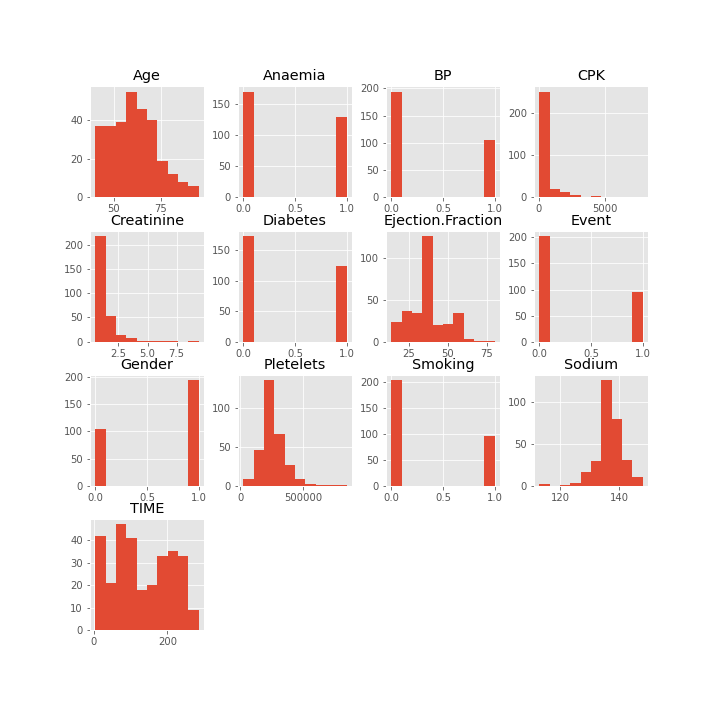
**Methodology:**

The dataset was obtained from the paper published recently: "Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* **20,** 16 (2020). https://doi.org/10.1186/s12911-020-1023-5".

Initially it was checked for missing values in the dataset, and was found to contain no null values. No new annotations of the dataset was altered and the entire dataset was used as is for subsequent analysis.

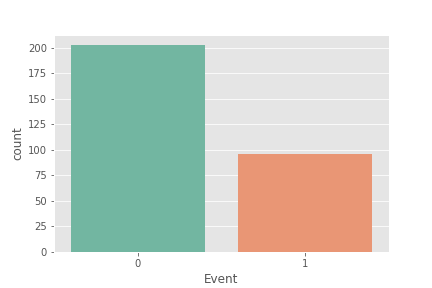
For data exploratory analysis, overall distribution of the dataset was computed and noted for each of the 13 attributes within the dataset. This distribution was done in tabular and graphical layout and the observations were made accordingly.

Next, correlation analysis was performed for each attribute against the target variable (Event) using pearsons correlation. Features, which corresponded to greater than 0.1 of correlation value were subsequently used as the set of features against which machine learning algorithms were to be tested. Hence, a smaller dataset only comprising of these relevant variables were extracted and then split into training and testing set (in the 0.7 and 0.3 ratio of the entire dataset respectively). Four different machine learning algorithms were then computed and ranked based on their accuracy.



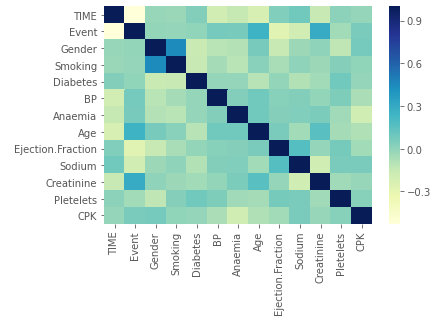
*Figure1: Counts distrubution of the dataset for all attributes of the study as mentioned in the previous section*

**Results:**

1) Descriptive analysis: Overall Heart failure prediction dataset was analysed for its sample distribution. Here, I observed that the dataset comprised mostly of older population (all patients were greater than 40 years of age) and the mean age was 60 years. Here, majority of the samples belonged to male population (roughly two-thirds of dataset). Again majority of patients in the dataset were survived patients (203 in number), while the dead patients (death event = 1) were only 96. In statistical terms, there are 32.11% positives and 67.89% negatives.

*Figure2: Death Occurences in patients due to Heart Failure. Red bar indicates death counts (96), while the blue bar represents survival counts (203).*

2) Correlation Analysis: A standard protocol of correlation of attributes to the target (i.e Event) was computed and plotted for the dataset using Pearson's correlation. None of the attributes had a strong correlation to the death of the patient, hence this problem may be a manifestation of multiple sets of features. Therefore, only those features whose absolute correlation value was > 0.1 were selected. They were Age, Ejection-Fraction, Creatinine, Sodium, and Time.



*Figure3: Correlation Matrix of the features among themselves using Pearsons Correlation.*

|  |  |
| --- | --- |
| **Data Attribute (Feature)** | **Pearsons Correlation Value** |
| Time | -0.52 |
| Creatinine | 0.294 |
| Age | 0.25 |
| Ejection Fraction | -0.268 |
| Sodium | -0.195 |

3) Machine Learning (ML) Model testing: A smaller dataset using relevant features (with absolute correlation values > 0.1) were derived and then this dataset was split up for training and testing sets ( 70% and 30% of the samples respectively). Four different ML approaches were tested and their F1-scores were computed. It was seen that K-Nearest Neighbor had the best accuracy compared to the rest.

|  |  |
| --- | --- |
| **Machine Learning Approach** | **Accuracy** |
| Linear Regression | 78.33% |
| Random Forest Classifier | 68.33% |
| K-Nearest Neighbor | 81.67% |
| Support Vector Machine | 75% |

**Observations:**

In this dataset comprising of older sample heart risk patients we could assess that the relation of 13 different features like hypertension (BP), Smoking, Age (>40 years), etc; were not independantly linked to the death of the patient. Howerver, all of these aspects could cumulatively be linked to the death of the patient. Here, we built and tested the efficacy of several machine learning approaches (K-Nearest Neighbor, Linear Regression, Random Forest Classifier, and Support Vector Machine). Out of these, K-Nearest Neighbor (KNN) gave the best performance of 81%. However it should be noted that the dataset was very small, hence this approach may not be the best one to describe heart failure in general, and thus further testing with more samples is advisable.

**Conclusions:**

KNN machine learning approach provides a better altenative for heart failure prediction in patients, when they were tested for the features comprising of age, Ejection-Fraction, Creatinine, Sodium, and Time.