Daniel Mejia

DSC-680

First Project - Heart Failure

**Business Problem**

By creating and system that can the detect the heart failure with high probability we can use the correct procedure to prevent, elongate or find a cure to the disease for the patient if it exists. In the medical field the detection of patient’s possibility of heart failure is important to prevent death. Also, the monetary compensation and saving a life would make the business profitable and the trademark should grow because of successful rate of preventing heart failure deaths. If the system can detect the possibility of heart failure better than doctors and prevent them, then our system is successful.

**Background/History**:

Heart failure is known to be when the heart muscles don’t pump blood well enough sustains the body required functions according to Mayo Clinic. Also, conditions like high blood pressure, viruses, arteries problems and related issues can affect the heart until it becomes weak and heart failure is imminent.

Exercising, reductions of salt, and managing stress can improve quality of life. But, if symptoms and conditions worsen, heart transplant and ventricular assist devices might be the next step to prevent death.

**Data Explanation**:

The data source will be Kraggle and for now one documents have been found further exploration for use and implementation is required. One is heart failure prediction Dataset. The file heartfailure.csv contains all the data to perform the analysis.

The data field were all in numerical type, this is optimal for AI and machine learning models. No categorical field were found in the document. No cleaning was needed as the document did not have blank field or missing data. This document was a high end cleaned filed.

The data set contains the following features in the heart failure file:

age – The age of the patient at time of condition (Integer)

anaemia – Decrease of red blood cells or hemoglobin (Boolean)

creatinine\_Phosphokinase – Level of the CPK enzyme in the blood, mcg/L (Double)

diabetes – Patient diabetes status (Boolean)

ejection\_fraction – Percentange of blood leaving the heart at each contraction, percentage (Double)

high\_blood\_pressure – Patient hypertension (Boolean)

platelets – Level of platelets in the blood, kiloplatelets/mL (Double)

serum\_creatinine – Level of Serum Creatine in the blood, mg/dL (Double)

serum\_sodium – Level of serum sodium in the blood, mEq/L (Double)

sex – Man status (Boolean)

smoking – Smokes status (Boolean)

time – Next appointment, number of days (Integer)

DEATH\_EVENT – Patient die before next appointment (Boolean)

**Methods**

Standard EDA will be used to explore and analyze the data to be implemented into a regression or another type of system to create a model for decision making. Cleaning of data will depend on the file data. Visualization techniques will be chosen to improve and find patterns to choose a model which might best fit the data representation.

**Analysis**

I started my analysis by using a correlation matrix map on the features. Found the fields to be most correlated to DEATH\_EVENT was time, serum\_sodium, age and ejection\_fraction.

Chart, scatter chart

Description automatically generated

Plotting the histogram for each feature allow the understanding and distribution of the data base on the graphical components. Out of the graphs a determination of left, right, or center skew were detected. The skew information allows different assumptions to be made about the specific data. A normal distribution graph allows for detection of anormal measuring.

A picture containing shoji, building, window

Description automatically generated

Box plots was used to identify where the data mostly lies and where data goes above the mean. The box plots allow for the data to be studied and bounds to be found easily. The graph below shows these limits for some of the data, data values were chosen based on their range to be place on the same graph for quick studying.

Chart, box and whisker chart

Description automatically generated

**Conclusion**

Alive and Death needed to be determined on a graphical level. The data was divided to show the points in different colors. On the scatter plots using different features from the data file. Grouping could not be determined to use KFNN model for the data. Also, linear model did not exist since the data was scatter all over, different plots and different features were tried to determine a pattern. The 3D map shows more people lived than died within the range of age 40 to 80. Age, ejection\_fraction, and platelets were used to determine the alive and death distribution using the 3D model. Also, just the age and ejection\_fraction was enough to determine distribution.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**Assumptions**

The assumption was in the pattern finding, I believe an easy pattern could be found if different variables were used to graph the data. However, changing the variables did not result in a pattern recognized by me with 2 years looking at different patterns in the data.

**Limitations**

The data limitation was on the age of the patient, patients started from 40 and forward until about 90. Data for younger people should be studied to understand if the death are proportional to the age or other factors.

Also, more data from the heart is required to understand the death of the patient which did not make it to their next appointment. Blood data is not enough to truly understand why their heart fail. Knowing the left ventricles, right ventricles functionalities and other related heart factors like the size is important to determine cause and improving accuracy of death probability.

**Challenges**

I haven’t created an AI system in a while, I am a bit rusty in creating this system that use words instead of numbers for analysis. Also, the probability of accuracy must be high when it comes to data that will affect the decision of doctors and nurses on a patient. Also, not enough data is in the document provided here and more data will be needed to create a more accurate system.

**Future Uses/Additional Applications**

The studied can be used to improve and create a better system where patients are filter better against the data found in their bodies. However, a better and more detailed research will need to be done. The accuracy could be improved with more data from the heart.

**Recommendations**

Data needs to be improved by partnering with health organizations and getting more data from these patients. Also, if the system finds the probability of a patient to be at risk of death truth, then shorter appointments should be done, to target and prevent death.

**Implementation Plan**

The system should be implemented in setting like a hospital, urgent care, or any other where data about the heart can be obtained and implemented into a pipeline for alert and prevention. This flow of information must be kept up to date. Also, the model should continuously be check for consistency.

**Ethical Assessment**

Personal data will be used to create this system and the system can be bias depending on skin color, race, and other factors. Consideration must be taken to not create a bias system

**Questions**:

1) How much will it cost to implement the system?

Implementing the system in a hospital setting to be real-time would require servers and devices to get the data automatically. The cost would be substantial but beneficial to the hospital.

2) How difficult is to implement the pipeline to the model?

The pipeline will require data to be fed to the servers and organized in a specific manner. However, once the pipeline is set up the process can automatically put the data digest and recommend patients base on probability of severity.

3) Reducing limitation on the model?

Implementing more data from the heart can reduce and increase the accuracy of detection, not limiting the system in accuracy. Also, the same system can be trained to detect other failures like kidney, liver, and similar issues base on the data.

4) What is the accuracy and if required how to increase it?

Just like above, getting more data for the required organ can improve the accuracy.

5) Is the business model being solve by the model?

The business model was solving to the specification of finding a model that detects heart failure better than doctor, in theory.

6) Scalability and maintenance requirements to keep it functional?

System must consistently be checked for consistency and data science must make sure new data does not deviate model from correctness.

7) How reliable is the system?

The system is only as reliable as the data, garbage in then garbage out.

8) How will the company make profit?

The company will profit from the recognition and surgeries to be done because of the detection system.

9) What are the benefits of implementing the system?

Benefits are substantial plus saving a life and being more accurate does bring fame and money.

10) Are there any issue with the model and how can they be fixed?

The model is limited by the features used and more testing must be done with new data and feature to find accuracy.

**Reference**:

*Heart failure - Symptoms and causes*. (2021, December 10). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142