Term End Milestone-1 (Project -1)

Predicting the Risk parameters for Cardiovascular patients

Proposal & Data Selection

DSC, Bellevue University

Supraja Rapuru

DSC680-T301 Applied Data Science (2225-1)

Professor Catie Williams

06/12/2022

ABSTRACT:

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.

Topic :

Predicting the Risk parameters for Cardiovascular patients

With this project I aim to probabilistically predict if a patient has a heart failure risk using machine learning models.

Business Problem :

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 and this dataset contains 12 features that can be used to predict mortality by heart failure.

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

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, hyperglycaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

Requirements, assumptions:

Requirements: This Project is trying to make use of machine learning techniques to automatically identify the risk factors. Here I am planning to feed the data of various risk factors to the ML model and the ML model will detect the high risk factors and send a signal to the prediction model to predict the mortality of the patient.

Assumption: Here I am assuming the Preprocessing operations were applied to the clinical data set available and made available data for feature extraction is not having any issue.

Costs and benefits:

Costs:

- The primary cost associated with this project is the time of the people working on it.
- Computing resources for modeling
- Data collection and processing computing costs

Benefits:

- Social benefit of this model is helping the patients to understand the high risk factors and take help of medical professionals in time
- Reduce the cost of medical costs
- To help avoid human biases

What Questions Are We Trying To Answer? :

- Which are the main predictor's features existing in the dataset identified?
- Which Model is going to fit our use case?
- Which is the best target variable for our model?
- What is the accuracy of the model?
- Do we got any other interesting facts from datasets? Like correlations etc
- Is it possible, the ML technique can identify the risk parameters predicting the mortality?

Datasets:

This project uses the UCI Machine Learning Repository's heart failure clinical records dataset

Describe Data.

Heart failure clinical records Data Set contains the medical records of 299 patients who had heart failure. The dataset contains 11 clinical features (some of them are binary, others are numerical), the follow-up period and the label DEATH_EVENT that indicates whether or not the patient has died.

We can find some features strictly related to medical aspects like levels of enzymes, sodium, creatinine and platelets in the blood and others that are more common like age, sex or smoking.

The table below lists out the details of all attributes in the dataset being used for the analysis

Attribute Name	Details about 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

Methods:

Classification models are a method of high importance used in various fields. In class determination, classification models are used to determine which class the data belongs to. The classification model is a model that works by making predictions. The purpose of the classification is to make use of the common characteristics of the data to parse the data in question.

As part of the regular classification model, I am planning to use some additional models as mentioned below.

Models were going to create by using Artificial Neural Network (ANN) and Deep Neural Network (DNN) algorithms for the feature dataset and by using the Convolutional Neural Network (CNN) algorithm for the image dataset, and classification processes were performed. Statistical results of sensitivity, specificity, prediction, F1 score, accuracy, false-positive rate, and false-negative rate were calculated using the confusion matrix values of the models and the results of each model were going to record in a table.

Ethical Considerations:

- This Data contains processed clinical records related to multiple anonymized patients and does not contains any PII-related information.

- Datasets and information on data were extracted from the public websites → UCL machine learning repositories.
- This data research is not going to harm any privacy.

Challenges/Issues :

Constraints: Major Constraints are related to used datasets , here the used datasets contain details about 299 patients and is highly imbalanced. It is a simulated dataset so the findings may not be scalable so we might not have to improve the effectiveness of dataset using SMOTE and other techniques.

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://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records)

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

D. Chicco, G. Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone", 2020

N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique, 2002

G. H. John, P. Langley. Estimating Continuous Distributions in Bayesian Classifiers, 1995