15CSE481

Machine Learning And Data Mining

- Case Study

Date: 04/11/2020

Team Details:

S.No	Name	Roll No
1.	Panakala Madhavi Bhavaghni	CB.EN.U4CSE17645
2.	SREEKISH MOOPIL	CB.EN.U4CSE17662
3.	SUNDEEP V V S AKELLA	CB.EN.U4CSE17664

Project Title: Heart Failure Prediction.

Faculty : Mrs. Archana R.

INDEX

Abstract

Cardiovascular disease is one of the major reasons for a person's death globally. Around 17.9 people die in a year due to heart failure. Cardiovascular disease can be prevented by addressing. behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol.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.

Introduction

A dataset which consist of details like sex, Level of CPK in blood is used to age is used to find out whether a person will suffer from heart failure or not. The Dataset is made up from the following informations such as age, Anemia (Boolean value), High Blood Pressure (Whether a person has hypertension or not), Creatinine phosphokinase (Level of the CPK enzyme in the blood), Diabetes(Bolean value), Ejection Fraction(Percentage of blood leaving), Sex (Binary value), Platelets (Platelets in the blood), Serum creatinine (Level of creatinine in the blood), Serum sodium(Level of sodium in the blood), Smoking (Boolean-If the patient smokes or not), Time (Follow-Up Period), Death Event(Target)

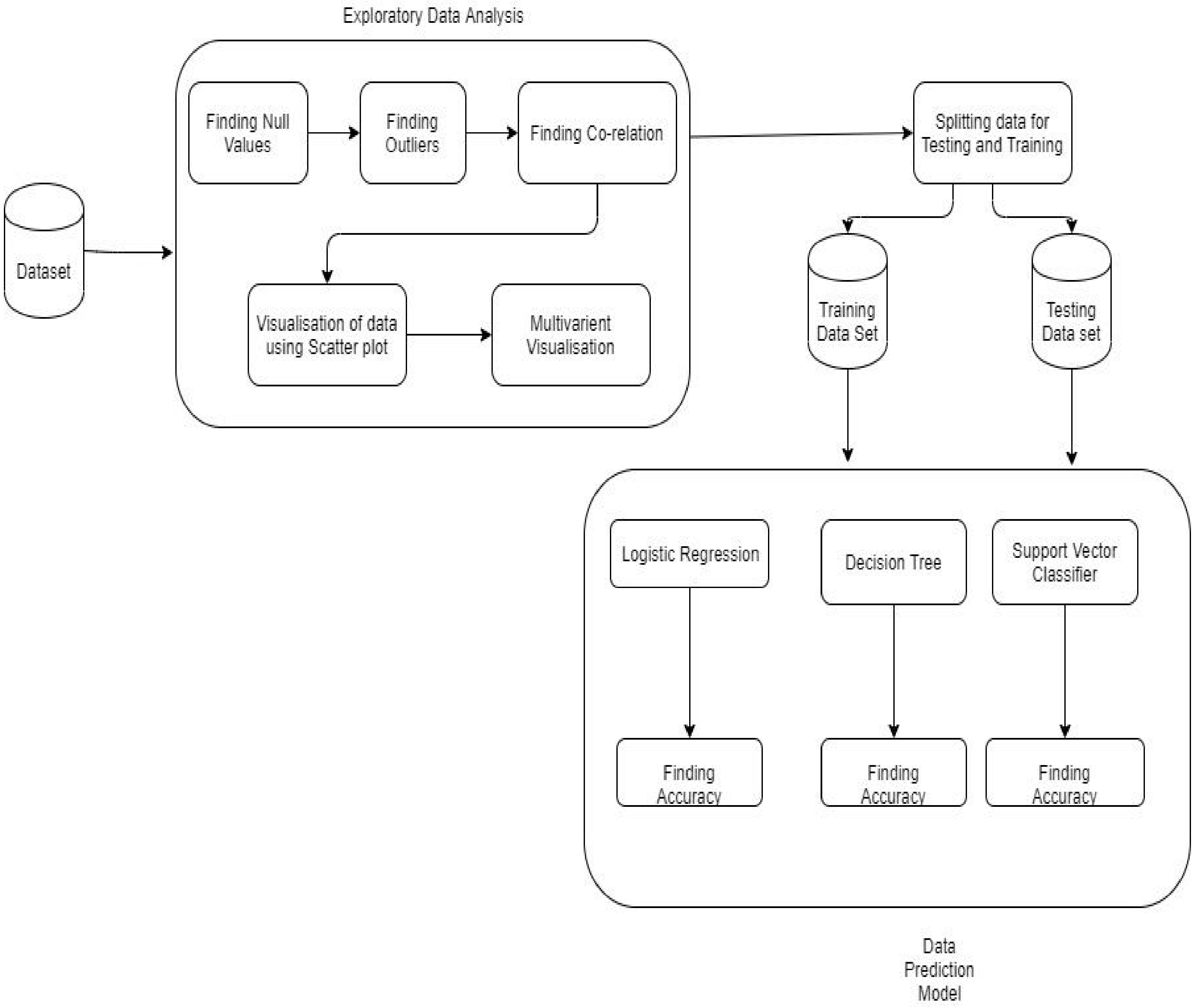
Objective

The main objective of the case study is to find out whether a person will suffer from cardiovascular disease or not. This is achieved by Performing a series of subtasks. The Subtasks Include finding Outliers , null values and correlation between each data. A brief analysis of data is also done. To learn more about the information in the dataset, data visualisation is also done. These above mentioned tasks come under exploratory data analysis. After getting a complete understanding of the data the model for predicting the target value can be built. The model is built using

Problem Statement

Death is inevitable. But trying to make a person live a little longer is possible.Physicians have been trying to predict heart attacks for as long as there have been heart attacks.Traditionally,they have relied on standard assessments of cholesterol,blood pressure,lifestyle factors and health conditions such as diabetes to predict whether a patient is likely to suffer a heart attack .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.People with cardiovascular disease or who are at high cardiovascular risk need early detection and management wherein a machine learning model can be of great help.

Architecture Diagram



Related Works And Novelty

The current ability to predict heart failure in patients is modest at best. It is unclear whether machine learning techniques that address higher dimensional, nonlinear relationships among variables would enhance prediction. We sought to compare the effectiveness of several machine learning algorithms for predicting heart failures.Methods and results: Using data that includes various factors to Improve Heart Failure prediction Outcomes ,we compared the effectiveness of different techniques.

Data collection and preparation

a. Feature Selection and/or Feature Extraction :

- During the feature selection,we have considered few references that are the main factors to cause the heart attacks/failures.Considering the other approaches and many other conditions and situations of a heart failure.We have considered few features for our case study

b. Data Visualization and Hypothesis:

- We have used many libraries like (matplotlib and seaborn) for plots and outliers to understand the behaviour and for the visualization of the data.

c. Data Cleaning & Pre-Processing:

- We have used the pandas library for the cleaning and the preprocessing of the whole dataset.Initially there were Nan values also,we cleaned the data by dropping such rows which didnt have a meaning to the other existing data

Data Modelling and Inference

Box plots and scatter plots are done.Also a correlation map is created and high correlation is observed between death and creatinine phosphokinase levels.

Conclusion

Three models which predicts CardioVascular Disease is built using the given dataset. A decision tree model with an accuracy of 0.883333333 is built for predicting CardioVascular Disease. A Logistic Regression model with an accuracy of 0.8833333 is also built using the dataset. A Support Vector Classifier model with an accuracy of 0.8833333 is also built. The accuracy for each model is found using the Confusion matrix.

Reference

https://towardsdatascience.com/exploratory-data-analysis-eda-python-87178e35b14

https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce

https://www.w3schools.com/python/showpython.asp?filename=demo_ml_scatterplot

https://scikit-learn.org/stable/modules/tree.html

https://realpython.com/logistic-regression-python/

https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-python/