Ashika Radhakrishnan

100761360

heart disease prediction

Contents

[EXECUTIVE SUMMARY 2](#_Toc55398852)

[PROBLEM STATEMENT 2](#_Toc55398853)

[RATIONAL STATEMENT 2](#_Toc55398854)

[DATA REQUIREMENTS 2](#_Toc55398855)

[ASSUMPTIONS 3](#_Toc55398856)

[CONSTRAINTS 4](#_Toc55398857)

[TEST PROCESS 4](#_Toc55398858)

# EXECUTIVE SUMMARY

This project is to predict if a patient has heart disease or not. The main goal is to analyze diseases based on various body parameters. Supervised learning approach would be good to create model for the dataset. I will also try to obtain some models such as SVM, K-means, Clustering, KNN and show how each variable behaves in each model.

Main goal of project is

* Classify if a patient has the disease or not.
* Create binomial classification model.
* Find the suitable variables for model and how variables behave in models.

# PROBLEM STATEMENT

Find out if a patient has heart disease or not form the provided data of various body parameters.

# RATIONAL STATEMENT

Binomial classification of the dataset and analyzing precision and accuracy. Applying all the classification models on dataset and find out confusion matrix.

# DATA REQUIREMENTS

* The first requirement is to find out the key variables which define the quality of the data.
* To create various models, the data should be large and diverse.
* The data should be in .csv or in excel format for better extraction.
* Get a dataset which a dependent variable on which another independent variable can rely on.
* The dataset should not have special characters in them so the data can be cleaned easily.
* The data is required to have numeric values.
* The data is required not to have null values and more outliers.
* The source is required to have more meaningful data in it.
* The data should be stored in properly for accessibility.
* The dataset is required to have a data description document, so everybody can understand the data better.

# ASSUMPTIONS

* Assume that the data is from a genuine source.
* No data manipulation was done before.
* We have enough data features available to build the machine learning models.
* For more model accuracy, we are assuming that the main dependent variable is highly related to all the independent variables.
* We have enough data to conduct analysis and it is normally distributed.
* All columns are necessary, and all independent variables are related to the dependent variable.
* We are going to use a classification predictive model to map from the independent variable to the dependent variable. We will use the classification algorithm to train the dataset and cluster it.
* We are using the binomial classification method.
* The dataset requires raw material.
* The model requires numeric values
* The dataset would be devoid of null values and the outliers.
* The data could be wrangled as per the model requirement for analysis purpose.
* The dataset would be easily understandable.

# CONSTRAINTS

* No proper information about data being balanced or imbalanced.
* No properly defined variables
* Presence of null values
* Data format should be in .csv or excel.
* As I am choosing binomial classification as a result, we are not getting any categorical classification of the analysis.
* If the outliers are useful or not
* Correlation of dependent variable with independent variables.
* If presence of string values in dataset

# TEST PROCESS

* Load aal the useful libraries
* Read data
* Perform Exploratory Data Analysis on dataset.
* Plot graphs and show relationship between variables.
* Remove null values
* Create Model for regression for various algorithms
* Find test accuracy for algorithms
* Calculate scores
* Compare models
* Create Confusion Matrix