|  |
| --- |
|  |
| Heart Disease Prediction |
| Statement of Work-V1 |

|  |
| --- |
| Bhavna Panwar  100802837 |

Contents

[Introduction 2](#_Toc381277520)

[Scope of Work 2](#_Toc381277523)

[Deliverables 2](#_Toc381277524)

[Milestones 2](#_Toc381277525)

[Gantt Chart 2](#_Toc381277523)

Data Source……………………………………………………………………………………………………………………………………………………………..2

Dataset Information…………………………………………………………………………………………………………………………………………………3

Data Assumptions…………………………………………………………………………………………………………………………………………………….4

Project Steps…………………………………………………………………………………………………………………………………………………………….4

[Testing 5](#_Toc381277529)

[Acceptance 5](#_Toc381277529)

**Introduction:**

*Cardiovascular disease or heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease. From WHO statistics every year 17.9 million dying from heart disease. The medical study says that human lifestyle is the main reason behind this heart problem. Apart from this there are many key factors which warns that the person may/may not getting chance of heart disease.*

Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications.

**Scope of Work:**

The objective of this project is to understand and analyze the factors that influence the heart disease. By using different machine learning regression models and comparing the results, we are going to achieve the best model which is more accurate and effective.

**Deliverables:**

At the end, there will be a python file and a document outlining the entire project that can be used to define variables that influence the heart disease and predicting the best effective model .

**Milestone:**

|  |  |
| --- | --- |
| Milestone | Estimated Delivery Date |
| Statement of Work | 11/06/2020 |
| Data Acquisition and Understanding | 11/23/2020 |
| Modelling | 11/23/2020 |
| Prototyping | 11/23/2020 |
| Deployment | 12/18/2020 |

**Gantt Chart:**

**Data Source:**

<https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression>

**Dataset Information:**

This dataset gives the information realated to heart disease. Dataset contain 13 columns, target is the class variable which is affected by other 12 columns. Here the aim is to classify the target variable to (disease\non disease) using different machine learning algorithm and findout which algorithm suitable for this dataset.

**Attribute Information**

• Sex: male or female(Nominal)  
• Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)  
Behavioral  
• Current Smoker: whether or not the patient is a current smoker (Nominal)  
• Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)  
Medical( history)  
• BP Meds: whether or not the patient was on blood pressure medication (Nominal)  
• Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)  
• Prevalent Hyp: whether or not the patient was hypertensive (Nominal)  
• Diabetes: whether or not the patient had diabetes (Nominal)  
Medical(current)  
• Tot Chol: total cholesterol level (Continuous)  
• Sys BP: systolic blood pressure (Continuous)  
• Dia BP: diastolic blood pressure (Continuous)  
• BMI: Body Mass Index (Continuous)  
• Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)  
• Glucose: glucose level (Continuous)  
Predict variable (desired target)  
• 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

**Dataset Assumptions:**

As the dataset is clean and every variable is important in our analysis so there are no assumptions required.

## **Steps to perform the classification**

The steps that will be followed to perform the classification are-

1. Import the dataset
2. Clean the dataset.
3. Data Visualization, in order to get the general idea about the data like the distribution of individual columns as well as determining the relationship between two variables
4. Preprocess the dataset. Preprocessing will include converting the string values into numeric format and converting the data which does not follow any ranking into One hot encoding. Also, preprocessing will include removal of outliers and scaling the data.
5. Model Fitting, various types of classification models will be fitted. Some of the models are-

* Logistic Regression
* Decision Tress
* Support Vector Machines
* Naive Bayes
* Random Forest Classifier
* Extreme Gradient Boost
* K-Nearest Neighbor

1. Model Selection, After the best model is selected, the next step will be moving towards model evaluation.
2. After the model has been trained on out training data, the model will be tested on testing data, which is separated from the training data. The model will be evaluated on various metrices such as accuracy, precision, recall, and f1-score. If the model is performing well on testing data, the model will be deployed on production, otherwise, one must rework on the data.

**Testing Process:**

**After the model has been trained on out training data, the model will be tested on testing data, which is separated from the training data. The model will be evaluated on various metrices such as accuracy, precision, recall, and f1-score. If the model is performing well on testing data, the model will be deployed on production, otherwise, one must rework on the data.**

**Acceptance:**

Date:06-11-2020

I Bhavna Panwar, a student at Durham College, consent to and accept the terms stated in this Statement of Work by initialing and signing each page below.

*(Durham College)*

By: Bhavna Panwar