**Statement of Work**

**Umutcan Asutlu**

**100797771**

**Background**

Heart diseases are significantly important topics in human health. Approximately, 12 million people die worldwide every year due to heart related problems. Early diagnosis can be crucial for saving human lives. With personal health data, artificial intelligence technologies can vastly help to reduce number of deaths caused by most common heart diseases like coronary artery disease.

**Objective**

The purpose of this project is predicting the patients’ risk of coronary heart disease in next 10 years. Patients personal health information will be used for data analysis methodologies and machine learning solutions. The project will detect people who are in high risk of having CHD in near future.

**Scope of Work**

With the help of the python this project will analyze the data that have been fed and will demonstrate insights to understand relations and answer questions about pattern of the disease. According to relations and insights, different machine learning algorithms and models with different hyperparameters will be tested and used to increase the accuracy which will help better to detect high risk to minimize the harm to patient in near future.

Although there are a lot of unknown things human body, it would still be insightful and useful to help people in advance.

**Data**

The required data is data of patients’ personal health information.

**Related Source**

The data is obtained from Kaggle (see reference 1). The data generated from Framingham Heart Study. It consists of 16 columns which are from left to right: 'male', 'age', 'education', 'currentSmoker', 'cigsPerDay', 'BPMeds', 'prevalentStroke', 'prevalentHyp', 'diabetes', 'totChol', 'sysBP','diaBP', 'BMI', 'heartRate', 'glucose', 'TenYearCHD'.

Sex: male or female (Nominal)

Age: Age of the patient;(Continuous)

Education: 1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = college

Current Smoker: whether 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.)

BP Meds: whether the patient was on blood pressure medication (Nominal)

Prevalent Stroke: whether the patient had previously had a stroke (Nominal)

Prevalent Hyp: whether the patient was hypertensive (Nominal)

Diabetes: whether the patient had diabetes (Nominal)

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.)

Glucose: glucose level (Continuous)

***Prediction Label***

10 Year CHD: Risk of coronary heart disease in 10 years (binary: “1”, means “Yes”, “0” means “No”)

It includes all attributes that are needed to make predictions and the label. There are total of 4238 entries in the raw data. There are a some of null values in data for certain properties. After removal there exists 3656 rows of data.

**Insights**

Age average of the patients is 49.5. Patients age range is 32 to 70.

Interestingly %49 of the patients are current smokers.

Risk of CHD is higher among lower educated people.

Age and systolic blood pressure highest positively correlated attributes with CHD.

While risk of CHD among female patients is %14, this number is %23.3 among male patients.

**Data Assumptions / Constraints**

The one big data concern is that there are not a lot of attributes about patients’ health vitals. They would increase the accuracy of the model. Not only that but also historical health data is generally crucial to make predictions since there is a chance to detect change and anomaly.

It is still a nice data contains sufficient amount of features to feed on machine learning algorithm.

The data can be supported with external data sources if needed. There may be other open data sources available online that matches with the existing data. However, there is no direct way to find additional information about existing patients.

**Tasks**

* Data needs to be understood well
* Data must be cleaned, all useless information must be removed, needed missing values must be handled. Exploratory data analysis must be done.
* Correlations and patterns must be defined. Relation maps must be generated.
* Different machine learning algorithms must be evaluated to pick best one to solve the problem.
* All algorithms should go through cross validation process to minimize variance
* Implement the evaluated model for prediction results.
* Design a cloud-based solution to host the python codes and the results.
* Deploy all python codes and machine learning models to the cloud solution to finish implementation
* Test and debug

**Deliverables**

All deliverables will be delivered through GitHub.

* Statement of Work
* Patients’ data
* Python code for data clean
* Python code that given input, generates visualization for data insights
* Python machine learning code makes prediction based on generated data
* A platform for using the project

**Schedule**

Statement of Work 11/06/2020

Understanding data

Data Acquisition and Gathering Results 12/1/2020

Development Process

Project Delivery 12/18/2020

**Testing Process**

The project will be capable of detecting high risk patients. Different algorithms with different hyperparameters will be used for both training and testing datasets to minimize bias and variance. To improve the accuracy the feature selection will be revised according to correlation matrix.

**References**

[1] <https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset/data>