Project proposal

***Data analysis on heart disease using the patient medical records.***

***Team details***: Project Group -1

|  |  |  |
| --- | --- | --- |
| S.no | Name | Net ID |
| 1 | Jayachandu Bandlamudi | bandlmd2 |
| 2 | Seunghyun Oh | soh45 |
| 3 | Augustine Chiu | achiu9 |
| 4 | Daulet Dyussekenov | dyussek2 |

***Motivation***: Health sciences is one of the area where Data analysis plays a crucial role and it offers a lot of medical records data for the patients. By analyzing the historical patient data, we can find some interesting insights about a certain disease, which will benefit in predicting a new patient getting a similar disease. So, that we can take necessary actions possible to prevent the disease or else we can treat the disease effectively by identifying the major cause.

For this project, we choose to work on chronic heart disease and it is one of major disease that causes thousands of deaths. As part of the Data analysis, we consider a sample of Male patient records from western cape, South Africa and through Data modelling on patient data we try to predict the chances of a patient getting chronic heart disease. Also, we figure out what are the major factors that cause heart disease, like

1. Is the heart disease depending on Age of the patient, then what is the age range that is more prone to the heart disease?
2. What effects does the Smoking and drinking habits have, and do they most likely cause heart disease?
3. How obesity or cholesterol, effects the chances of getting a heart disease?
4. Does the heart disease depended on family history? i.e; If any family member has the heart disease then how likely it is, patient getting a heart disease.

Since task is predicting if the patient has chronic heart disease which is a label (1/0) 1 – if the patient has the disease and 0 otherwise. So, we can treat the problem as a binary classification task and we may use several supervised learning models for the analysis.

**Data Description**:

Source URL: <http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data>

Data dictionary: <http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.info.txt>

For this task our target/response is “**chd**”, which specifies whether a patient has chronic heart disease and we have 9 independent/predictor variables using which we do predictive modelling for the “**chd**”. These 9 predictor variables as follows ‘**sbp**’ (systolic blood pressure), ‘**tobacco**’ (cumulative tobacco in kg), ‘**ldl**’ (low density lipoprotein cholesterol), ‘**adiposity**’, ‘**famhist**’ (family history of heart disease), ‘**typea**’ (Personality trait that is characterized by excessive competitiveness and aggression) , ‘**obesity**’, ‘**alcohol’**, ‘**age**’ . Among the 9 variables ‘**famhist**’ is categorical with two levels(Present/Absent) and the remaining variables are continuous/ numerical data type. Also, we have **462** patient records as observations for the analysis.

**Data Analysis intended to perform**:

We intend to partition the whole data into training dataset with 90 percent of data, validation dataset with 10 percent of data. And training dataset is used for building predictive model, where-as validation dataset will be used for validating model performance. Along with the predictive modelling, we try to answer the significance of several variables by exploring/ diagnosing variables and variable interactions, transformations w.r.to response.

***Task assignment*:**

Daulet – Considering **‘chd’** as response, since it has two levels (0/1) so we can use logistic regression to predict what are the odds of the patient having heart disease depending on significant predictors.

Jayachandu – Considering the Combination of **‘chd’, ‘famhist’** we create a new response variable(**‘Class’**), with several classes such as in the below table

|  |  |  |
| --- | --- | --- |
| Chd(0/1) | famhist(Present/Absent) | Class (New response) |
| 0 | Present | Class1 |
| 1 | Present | Class2 |
| 0 | Absent | Class3 |
| 1 | Absent | Class4 |

Based on the new response, which has 4 classes so we use discriminant analysis (Linear/ Quadratic) to predict the ‘**Class**’. And using this analysis we try to answer how the combination of **‘chd’, ‘famhist’** variables effected by the several other continuous predictor variables in the data.

Augustine – In the data we have several variables, and **‘sbp’** is one of them. Here we try to answer whether a patient with certain **‘sbp’** (blood pressure level) tend to have heart disease more likely. So, we would perform ANOVA for **‘sbp’** based on **‘chd’, ‘famhist’** and **‘cluster\_group’**. where **‘cluster\_group’** is the new categorical variable which will be assigned by performing ‘Clustering’ on the continuous variables in the data other than **‘sbp’.** By doing clustering we group the similar patients into one group/cluster and the optimal number of clusters will be decided based on reasoning and analysis

Seunghyun – Variable ‘**typea**’ describes the personality of patient such as if he is aggressive/impatient then his ‘**typea**’ score is high. So, here we try to answer whether ‘typea’ score has association with rest of the continuous variables related to patient habits (Alcohol, Smoking, Eating) using Generalized Linear model with ‘**typea**’ as response and remaining continuous variables as predictors.

\*\* \* THE END \*\*\*