**Exercise #1: Using Machine Learning Agorithm to predict Heart Disease**

**Group:**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Name** | **Student ID** | **Responsibility** |
| **1** | **The Cuong, Bui** | **C0770339** |  |
| **2** | **Chiranjeev Singh** | **C0760944** |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Introduction**

**Machine Learning Introduction**

**Process of Building Machine Learning Model**

**Programming Language and tools used for excercise**

**Understanding Features and data sets of heart diseases.**

**Implementation Machine learning algorithm using Python for predicting Heart Diseases**

**Introduction**

Heart diseases reflects a various conditions in which can affect human’s heart.

Diseases relates to the heart disease are various include coronary artery disease (blood vessel diseases), arrhythmias (heart rhythm problems) and congenital heart defects (heart defects you're born with) and so on..

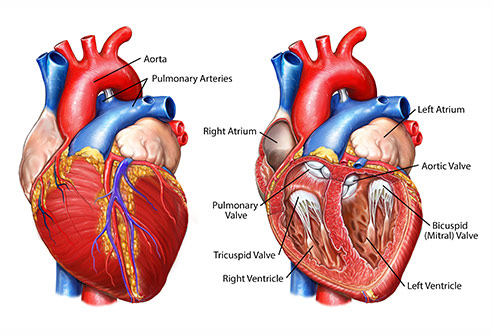


Figure 1: Heart parts

Heart diseases have been considered as the serious causes of high death rate of human life. Hence scientist has been spent a lots of effort in researching the effective methods to quickly diagnostic and prevent heart disease and it has become more than necessary. With a huge data from large number of hospitals and clinics in all over the world, it can help to build a good data-driven systems to analyze and predict heart diseases and then can make the research and prevention process more effective. From that it can ensure that more people can live a healthy life.

With the development and maturity of Big Data along with new technologies such as Cloud Computing, nowadays, Machine Learning is attracting data scientists apply across many spheres around the world. The healthcare industry is one of the realistic cases in which Machine Learning can play an essential role in predicting presence of Locomotor disorders, Heart diseases and so on. And machine learning the predictions made by Machine Learning are quite accurate and promising coming more accuracy. With the maturity and increasing experience of Data Scientist in applying high technology in researching heart diseases, it is promising that Machine Learning can provide a pivotal insights to hospital or doctors to adapt their diagnosis and treatment method for each individual patient.

Regarding to above reasons, our group are pleased to apply Machine Learning to predict heart disease as an exercise of Big Data Fundamental course of Fall 2019 at Lambton College. This is not only to help us enrich knowledge, experience in Machine Learning as the objective of the course but also help us to have a right feeling that we are going to the right course in which we can participate in the most top health problem of the world in which we can contribute our knowledge to insights for society.

Machine Learning Introduction

Building Machine Learning process Process

## 

## Iterative Machine Learning Process

## 

## Data Preparation: How to import raw data, what are the most cleaning data methods.

## 

## Feature Engineering: How do you turn your raw data to relevant data, meaningful for machine learning algorithm include: informative, discriminative, non-redundant. It results in improved model performance on unseen data.

## Feature engineering usually includes:

## Feature construction:

## Feature Transformation: is the process of transforming a feature into new one with a specific function.

## Example: Scaling (the most important), log (reduce heteroscedasticity)

## Dimension reduction (selection and extraction): reduce number of irrelevant features used to build the models with the goal of keeping only informative, discriminative non-redundant features. It results to the advantage of faster computation, less storage space, improve model performance, data visualization.

## . How can you make the difference between useful and useless data in huge dataset.

## Data Modeling: What are different types of Machine Learning Algorithm, Choose the most appropriate one to build model reflect to dataset.

## Train model using dataset using a learning algorithm

## 

## Performance Measure: What is the right method to access the performance ? Which indicator is used?

## Peformance Improvement:

## Reason the performance of built model is not well.

## Most common techniques to improve performance

## Tools to be used:

## Programming Language: Among Python, R and other programming languages, group decided to use Python for our exercise as it can help us learn practical language for our future work.

## Jupiter: a popular interface for Machine Learning. Using notebook format with input cells contains code and output cells contains the results of code.

## Advantage: it can iterative code quickly, instantly visualizing the result of modification.

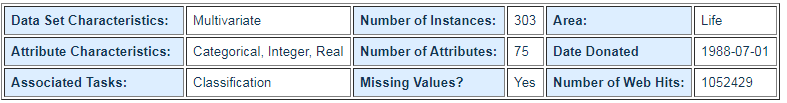
## Pandas: is the reference module to efficiently manipulate millions rows of data in Python.

## Scikit learn: is one of the reference modules for Machine Learning in Python.

## Numpy and SciPy and Matplotlib are more convenient modules for data computerization and data visualization

**Understanding Data sets of Heart Diseases:**

For this exercise, we used the most recommended data set called Cleveland Heart Disease from a study of heart disease from UCI Machine Learning Repository which is maintained by the Center for Machine Learning and Intellignet Systems of the University of California, Irvine.



In particular, the Cleveland data set is the only one that has been used by Machine Learning researchers to this date. The main target of most of the research is to detect the presence/absence of heart disease from patient’s data. Experiments with the Cleveland data sets have been concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

For exercise purpose and initial start, our group select the “processed” of Cleveland data to ensure higher accuracy of prediction.

The dataset consists of 76 attributes and 303 individual data. However, all published experiments refer to using a subset of only 14 of them, which are described below.

1. *Age* : the most important risk factor in developing cardiovascular or heart diseases. It is estimated that the high risk of coronary heart disease and stroke for people over 55 or 65 and older.
2. *Sex* : 1 = male, 0 = female

Men is considered greater risk of heart disease than pre-menopausal women. However, If a female has diabetes, the higher possibility to develop heart disease than a male.

1. Angina *(Chest-pain type)* : when heart muscle doesn’t get enough oxygen-rich blood, it causes Chest pain or discomfort.

1 = typical angina, 2 = atypical angina, 3 = non — anginal pain, 4 = asymptotic

1. *Resting Blood Pressure* (in mmHg (unit)): high blood pressure can damage arteries that feed heart over time.
2. *Serum Cholestrol (* mg/dl (unit)): A high level of low-density lipoprotein (LDL) (“good”) cholesterol is most likely to narrow arteries, a high level of high-density lipoprotein (HDL) cholesterol (“good”) lowers your risk of heart attack.
3. *Fasting Blood Sugar* : blood sugar levels in human’s body can rise up if it is not produced enough of a hormone secreted by body’s pancreas (insulin). If fasting blood sugar > 120mg/dl then : 1 (true), else : 0 (false)
4. *Resting ECG(*electrocardiographic): 0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hyperthrophy
5. *Max heart rate achieved*: The increase in the cardiovascular risk, associated with the acceleration of heart rate.
6. *Exercise induced angina* : The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. 1 = yes, 0 = no
7. *ST depression induced by exercise relative to rest* : displays the value which is integer or float.
8. *Peak exercise ST segment* : A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after the J point. 1 = upsloping, 2 = flat, 3 = downsloping
9. *Number of major vessels (0–3) colored by flourosopy*.
10. *Thal* (thalassemia): 3 = normal,6 = fixed defect, 7 = reversible defect
11. *Diagnosis of heart disease*: 0 = absence, 1, 2, 3, 4 = present.