**INFOSYS 722**

Tapas Mishra (student ID = 341236064)

University of Auckland

Master of Professional Studies – Data Science

**Iteration 3 - BDAS**

1. **Business/Situation understanding –** Identify potential future heart patients based on various health parameters for targeted preventive and wellness programs.

## 1.**1 Identify the objectives of the business/situation**

The debate over health care reform and the future of the Affordable Care dominated headlines in 2017. The onus is on major healthcare insurance providers to bring their cost down , intern reducing the premium charged on members, making health care more affordable. Biggest question is HOW ?

**One way could be, keep your members (insurance subscribers) healthy !**

Since there is a worldwide rise in the prevalence of diseases, preventive healthcare or Members Wellness programs could be one of the services that should be given to members in the customer relationship management concept. Healthcare insurance providers can utilize the huge member’s lab data to identify the segments of members , who needs preventative healthcare.

Reduction in insurance premium for subscribers.

Target Investment in members Wellness Programs.

Healthcare more affordable.

Reduction in insurance claims (Cost for insurance provider)

Healthy members

* There are many **Lifestyle diseases** such as atherosclerosis, heart disease, and stroke; obesity and type 2 diabetes; and diseases associated with smoking and alcohol and drug abuse. Regular physical activity helps prevent obesity, heart disease, hypertension, diabetes, colon cancer, and premature mortality.
* **In our study , we will focus on heart diseases as a proof of concept , which can be later scaled to other diseases.**
* With effective patient data tracking and data analytics, patients with potential diseases could be identified. Also, better life choice practices could be recommended which would prolong their life expectancy.

## 1.2 **Assess the Situation**

* **Current Trend :** Nowadays healthcare industry generates large amount of data about patients, disease diagnosis etc. Data mining provides a set of techniques to discover hidden patterns from data. A major challenge facing Healthcare industry is quality of service. Quality of service implies diagnosing disease correctly & provides effective treatments to patients. Poor diagnosis can lead to disastrous consequences which are unacceptable.
* **Data :** Since this is an independent study , the datafor this experiment will be downloaded from UCI Machine Learning repository. There is no data purchased from any external vendor. Although there is plenty of data on multiple diseases, we will restrict our analysis to patients with heart disease. If successful , the program can be expanded.
* **Computing Capabilities :** This study will be performed on machine with 1TB data storage , 8GB RAM , Intel core i5 processor , and 64 bit Windows 10 Operating System. This system configuration is enough for the study as our dataset is of few MB’s and is also suitable for our data mining tools.
* **Team :** It's clear that, we have in-house expertise in database, data warehousing and data cleaning but little experience in statistical analysis. Thus statistical expert may be consulted. We may also need to consult to any medical / Heart specialist SME, to understand trivial health parameters and their relationships.
* **Risk :** Aside from time spent by analyst on the study, there is not a great deal of immediate risk in this venture. However, time is always important, so this initial project is scheduled for a single university semester. In addition , as suggested by experts , business scope can be

## 1.**3 Determine data mining objectives**

Using appropriate datamining model and visualization technique we want to achieve below objectives

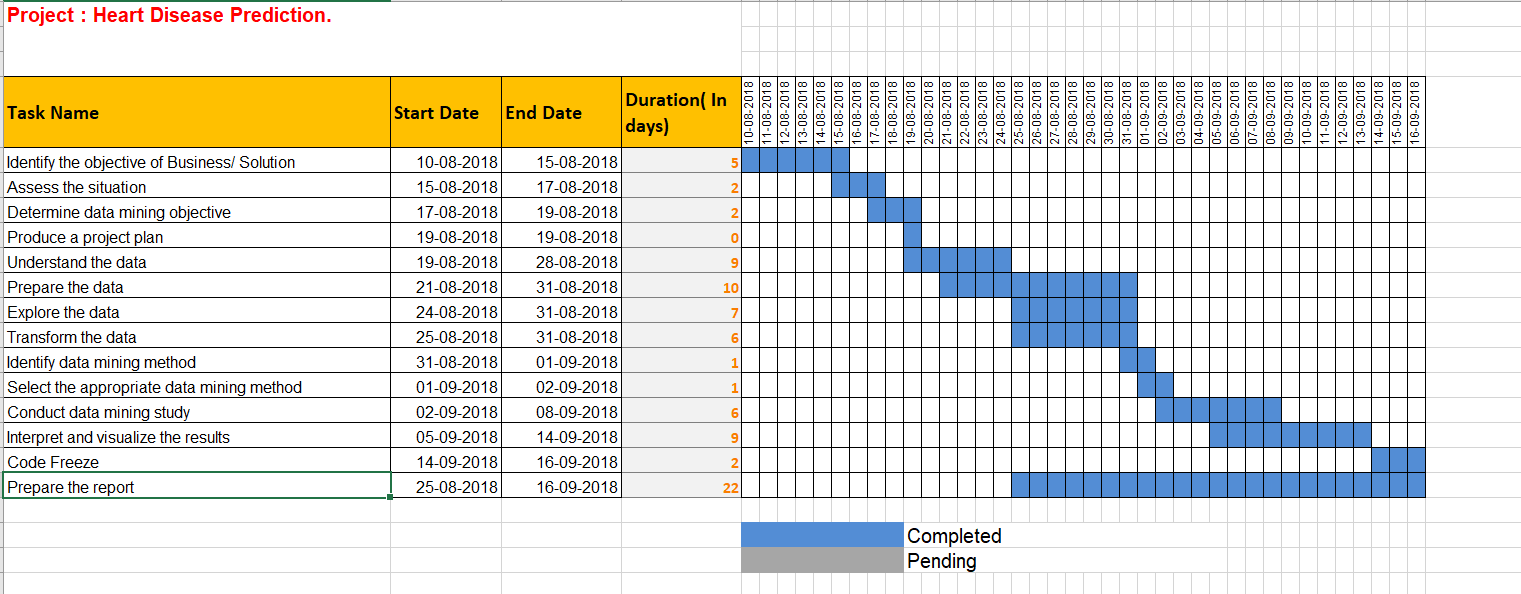
* Predict future potential heart patients based on their current health parameters.
* Identify top n health parameters , which are critical indicators of developing heart disease in future.
* Create a datamining model , which can use patient lab data and can consistently predict potential heart patients , so that those can be used for targeted wellness programs.

In addition, different health promotion and training services will be recommended to such subscribers.

This is a Supervised Learning project and can be solved using various data mining techniques, however we will focus on Classification type problem to level set the scope of the project.

## **1.4 Produce a Project plan**

* The project follows a sequence of steps that is a synthesis of the Cross-Industry Standard Process for Data Mining the KDD process (Fayyad et al., 1996).
* Below chart shows , how we have segregated different phases of the project and defined their timelines.
* Tasks such as “Prepare the data” and “Conduct data mining” activities are given maximum time , because they are more of a iterative process.
* Below chart shows most of the task as completed status , because the snapshot is post completion of the project.



# **2. Data understanding**

## **2.1 Collect Initial Data**

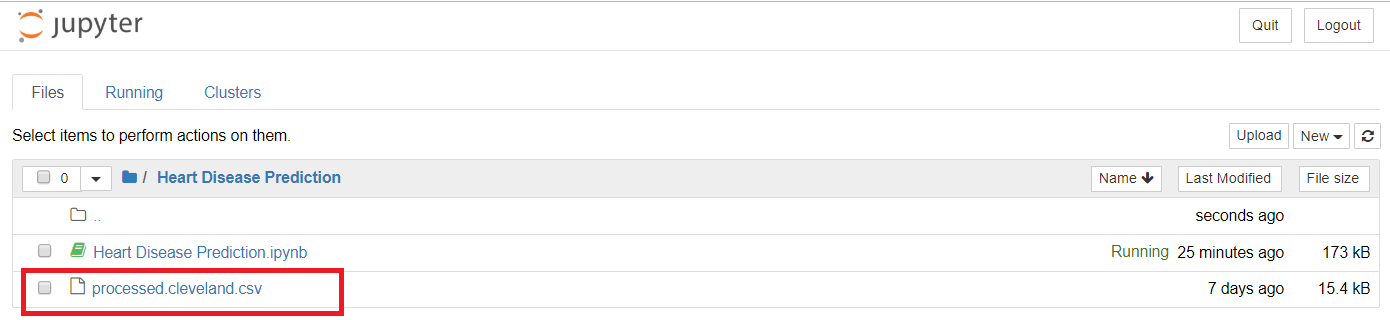
* The initial data was collect from <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>
* There was no data purchased from any external source.
* No prior registration is required by the user to access the dataset.
* The data was available in the form of **csv**(Comma Separated Value).
* There were **no issues** encountered while collecting the data as the data is hosted by the UCI machine learning and is available for free of cost
* **If in future any more data is required for analysis, it will be included for the analysis suitably.**

**Note: The data does not have any copyrights and can be used by any individual. One need not have any prior permissions either to access the data.**

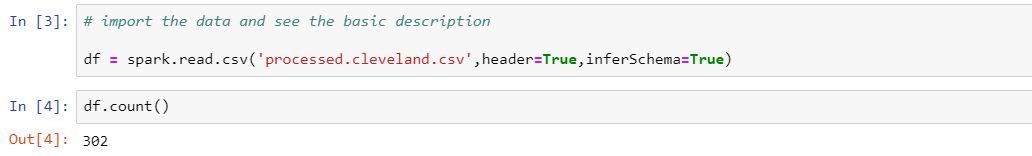
## **2.2 Describe the Data**

The dataset used in this project is part of a database contains 14 features from Cleveland Clinic Foundation for heart disease [1]. The dataset shows different levels of heart disease presence from 1 to 4 and 0 for the absence of the disease. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0). We have 303 rows of people data with 13 continuous observation of different symptoms.

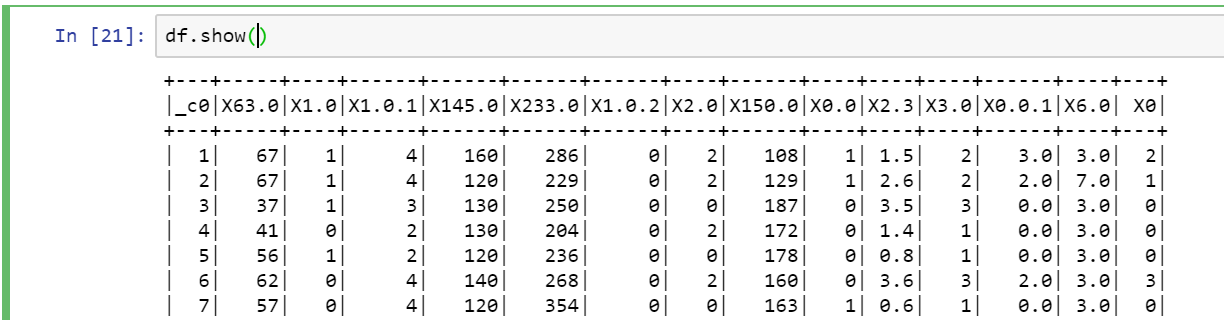
After the dataset was collected, it was uploaded to the ./Heart Disease Prediction folder and analysed.



* The value of the data is only numeric
* Total number of rows = 303

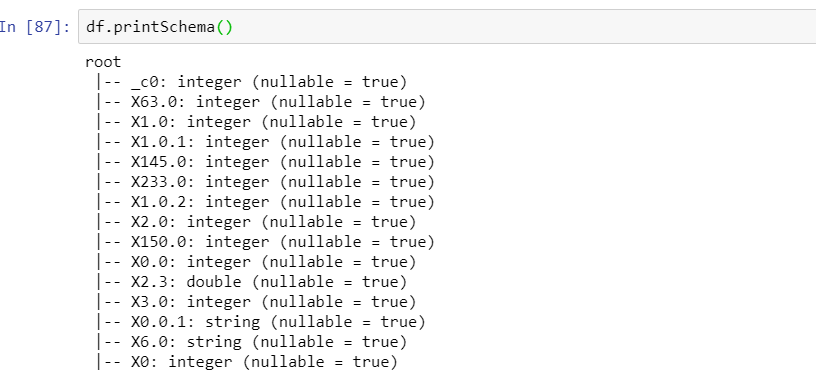


* Data contains 14 features , which are different health parameter of patients. All the features , will be taken into consideration.
* Here is the snapshot of the data exported from the data source , with some initial rows, although it does not have column names , which we are going to define in the later stages based on the data dictionary.
* By , just looking at data snapshot , we can see some categorical type values , however they are already encoded. So we may not be required to encode string to numeric values for these variables.



By , just looking at data snapshot , we can see some categorical type values , however they are already encoded. So we may not be required to encode string to numeric values for these variables.

* Most of the datatypes are integer and double , however few of them are string datatype, although they contain numeric values. **That shows , there could be some missing or NA or default string type values.** We will explore them and decide to impute.



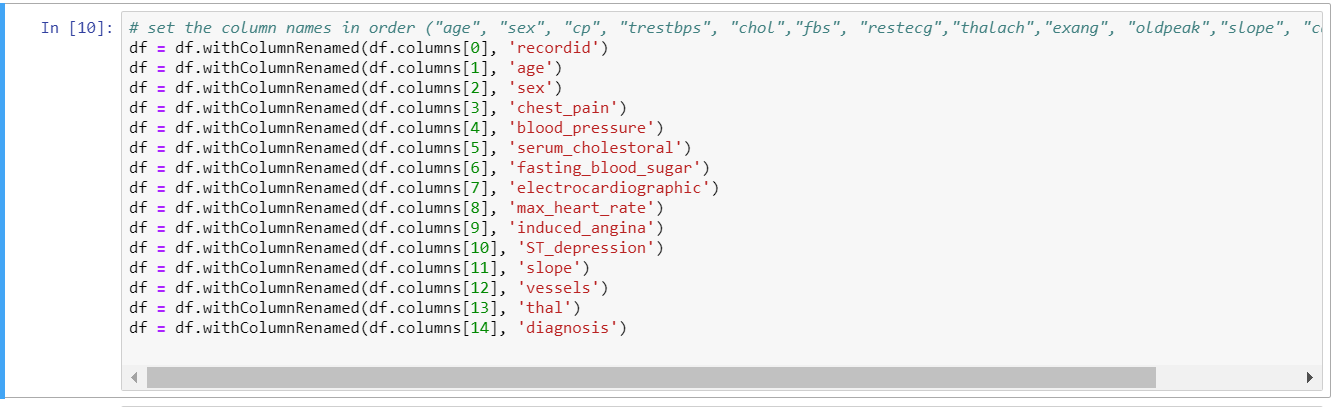
Below are the understand of each attributes and type of values they contain.

|  |  |  |  |
| --- | --- | --- | --- |
| **S no** | **Parameters** | **Parameter description** | **Values** |
| 1 | age | Age in years | Continuous |
| 2 | sex | Male or female | 1= male , 0= female |
| 3 | thestbps | Resting blood pressure | Continuous value in mmHg |
| 4 | cp | Chest pain type | 1= typical type 1 2= typical type angina 3= non-angina pain 4= asymptomatic |
| 5 | chol | Serum cholesterol | Continuous value in mm/dL |
| 6 | fbs | Fasting blood sugar | 1≥120 mg/dL 0≤120 mg/dL |
| 7 | restecg | Resting electrographic results | 0= normal 1= having ST-T wave abnormal 2= left ventricular hypertrophy |
| 8 | thalach | Maximum heart rate achieved | Continuous value |
| 9 | old peak | ST depression induced by exercise relative to rest | Continuous value |
| 10 | exang | Exercise induced angina | 0= no 1= yes |
| 11 | Ca | Number of major vessels colored by fluoroscopy | 0–3 value |
| 12 | slope | Slope of the peak exercise ST segment | 1= unsloping 2= flat 3= downsloping |
| 13 | thal | Defect type | 3= normal 6= fixed 7= reversible defect |
| 14 | obes | Obesity | 1= yes , 0= no |
| 15 | num | Diagnosis of heart disease | 1 to 4 , angiographic disease status depending on diameter narrowing. |

## **2.3 Explore the data**

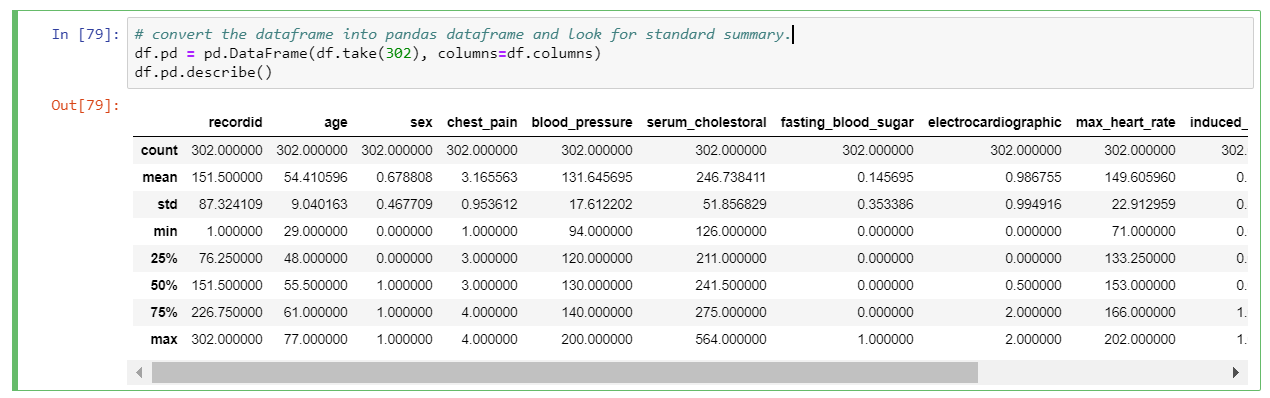
**Iteration 1 : In iteration 1 of data exploration , I will look at basic statistics of my dataset and explore various parameters , that predominate in patients with heart disease.**

Let us first give the column names to dataset , so that they can be easily understood.



Now, I first convert spark dataframe into Pandas dataframe , because it is easier to explore and manipulate data frame in pandas.

I first start to look at the basic statistics of attributes in the dataset, using describe() functions of pandas.



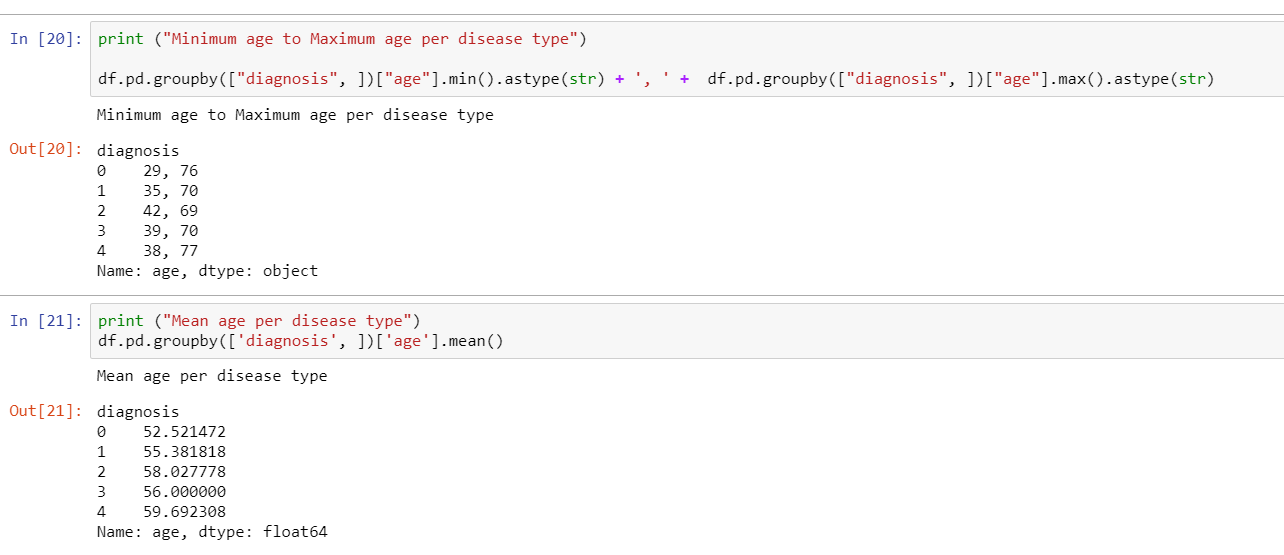
Dataset has mean(age) of 54, with min of 29 , so it shows , dataset is more skewed towards adults and old aged people , that is common understanding, teenagers don’t have heart disease in common.

However, I do not get detailed understanding of the my data by basic statistics, I will have to explore each feature.

### **Let's find the ranges of each feature by disease type, to find out what are some *expected* range of health parameters to have a heart disease in patient.**

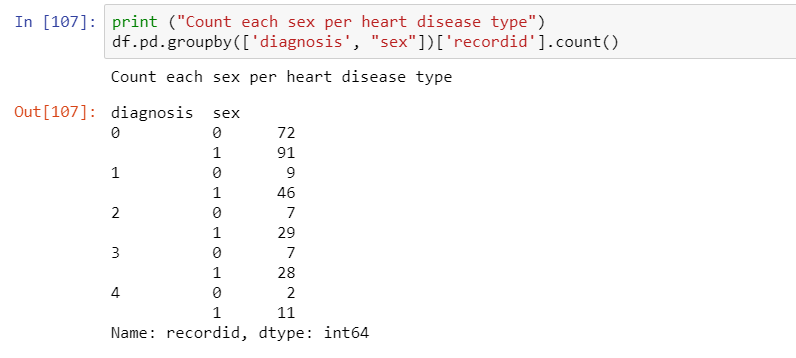
1. **Age :**

#### We can see that heart disease with type 3 can be present in lower age then type 2



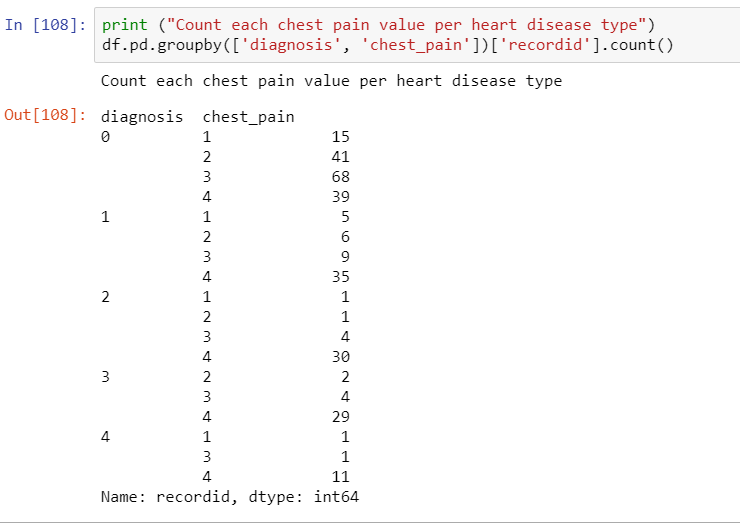
1. **Sex :**

#### We can see that heart disease all types can be present in men with higher probability than in women

****

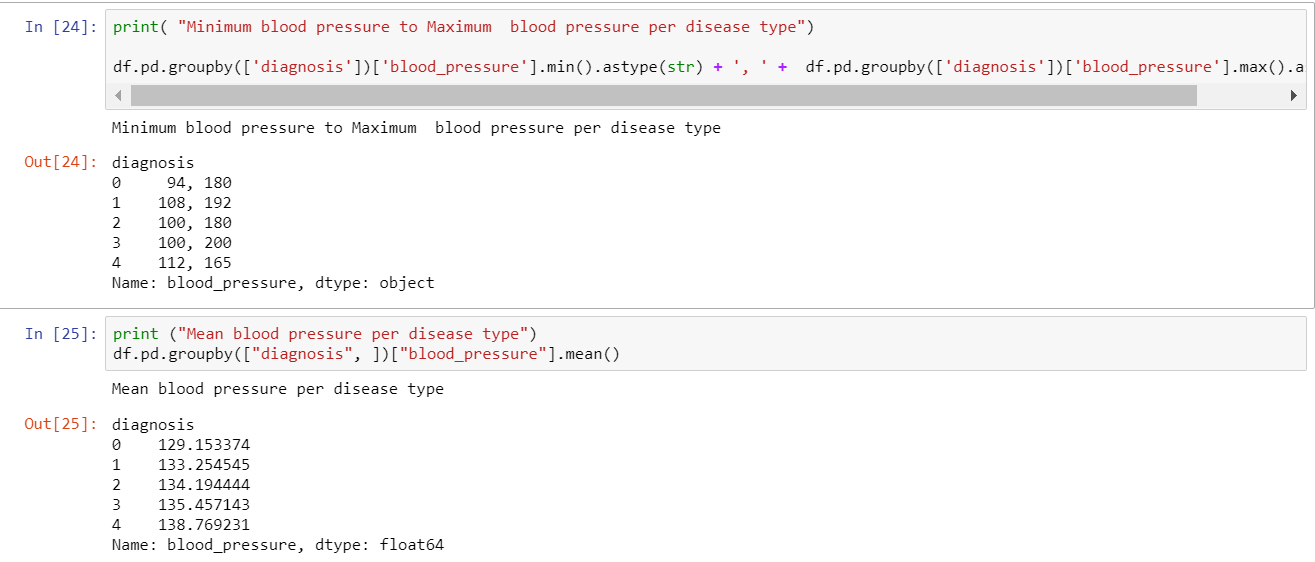
1. **chest\_pain:**

#### The people with chest pain = 4 often have higher chances of heart disease.

****

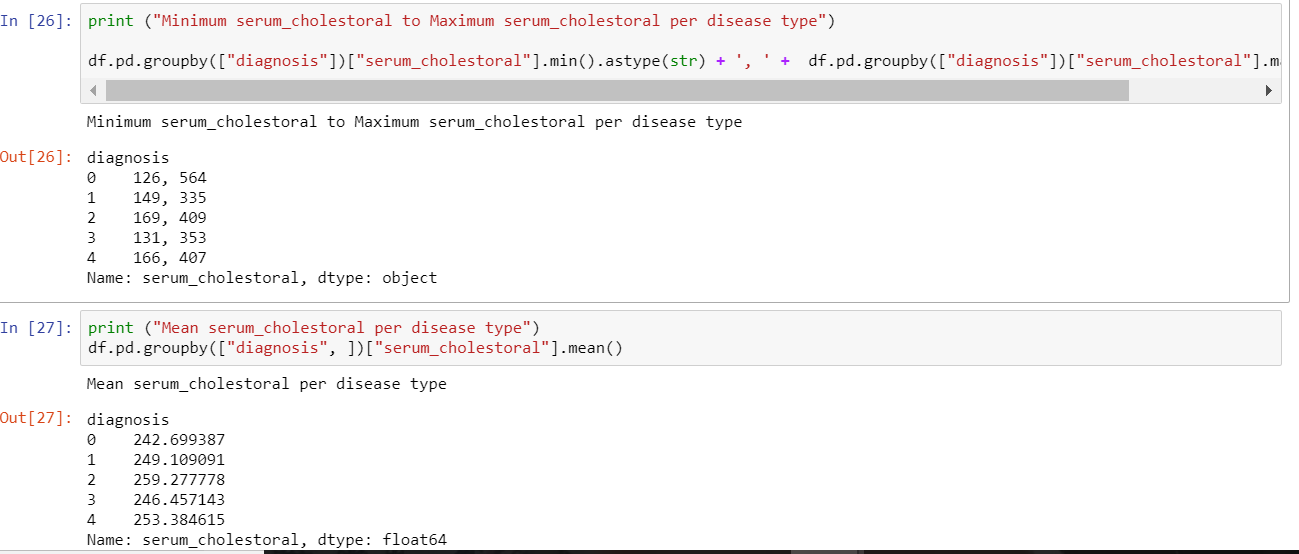
1. **blood\_pressure:**

#### As bigger is mean blood pressure as higher is type of heart disease

****

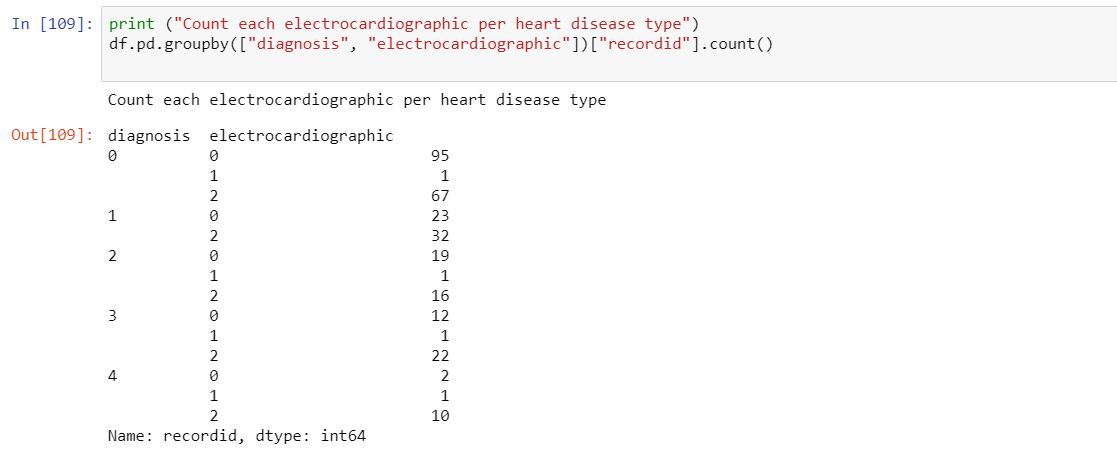
1. **serum\_cholestoral:**

*We do not get any clear direction from both range and mean cholesterol values, they are spread almost each disease type. It does not looks to have good predictive importance.*

****

1. **electrocardiographic:**

*patients with electrocardiographic 2 , generally have higher type of heart disease.*

****

1. **Similarly , I explored max\_heart\_rate, induced\_angina, ST\_depression, slope, vessels, thal, fasting\_blood\_sugar**

*(Screenshot not attached , but available in code)*

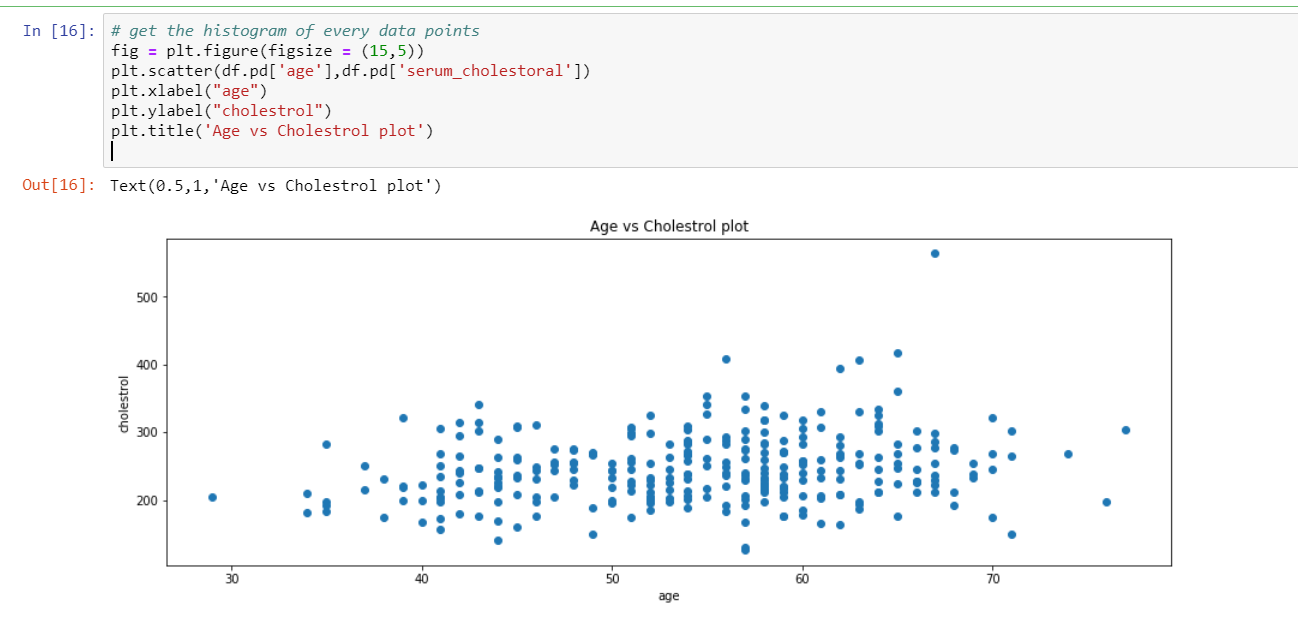
By above analysis , I can say person with

* age > 38
* man
* with chest pain = 4
* blood pressure > 112
* fasting\_blood\_sugar = 0
* electrocardiographic = 2
* max\_heart\_rate > 114
* ST\_depression about 2
* slope >=2
* vessels about 1.6
* thal more than 6

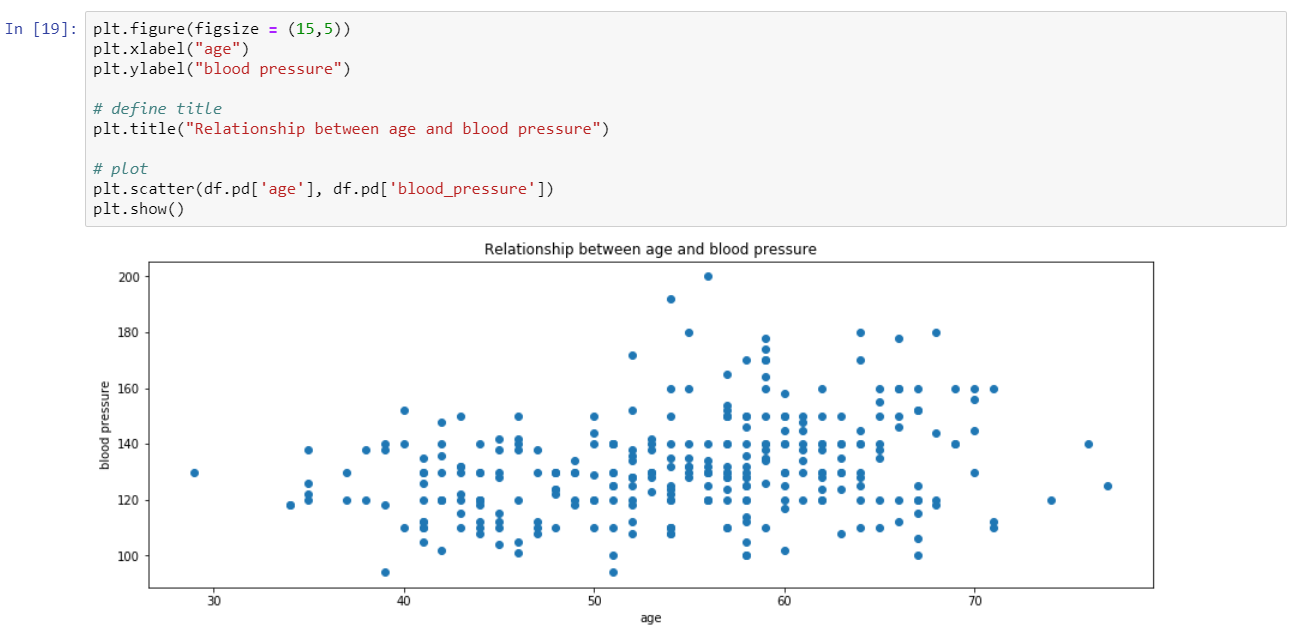
**is the most likely to have 4 type of the heart disease**

**Iteration 2 : In this iteration , I will try to understand relationships , between several feature using some visualization.**

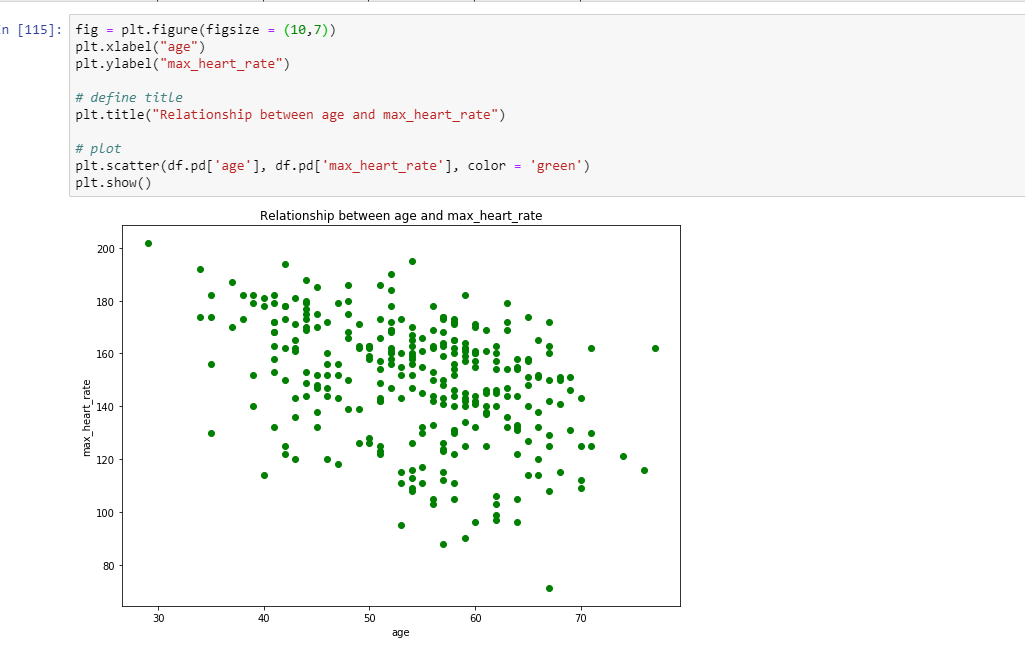
1. **Age Vs Cholesterol :** I can clearly see, that , number of people with higher age have higher cholesterol. There seems to be 1 instance at age=68, who have cholesterol = 580. It looks to be a outlier in our data.



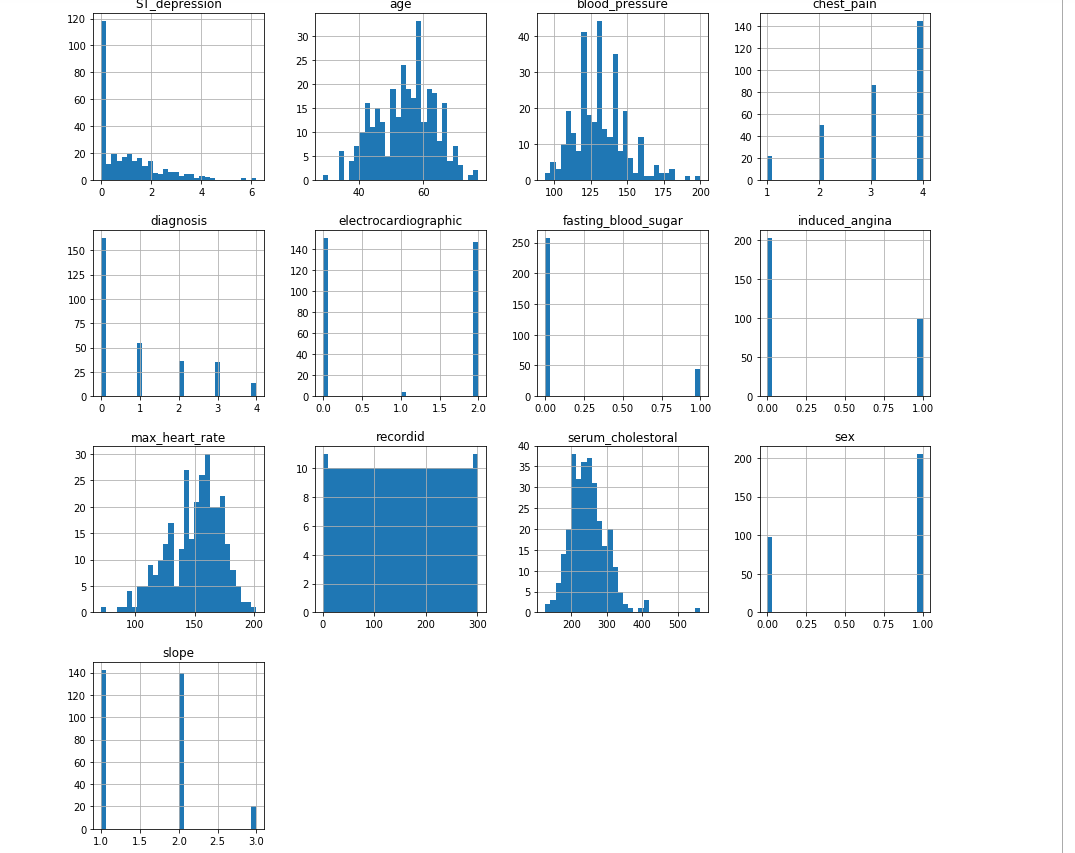
1. **Age Vs Blood Pressure :** I can see , that many people beyond age of 55 have higher blood pressure as compared to people of age below 55.



1. **Age Vs Maximum heart rate :** I see there is negative correlation between , age and max\_heart\_rate. Maximum heart rate reduces with age.



1. **Histogram of data** : With simple histogram of our data, I can easily observe the distribution of different attributes. One thing to note here is the fact that it is extremely easy for us to see which attributes are categorical values and which are not.



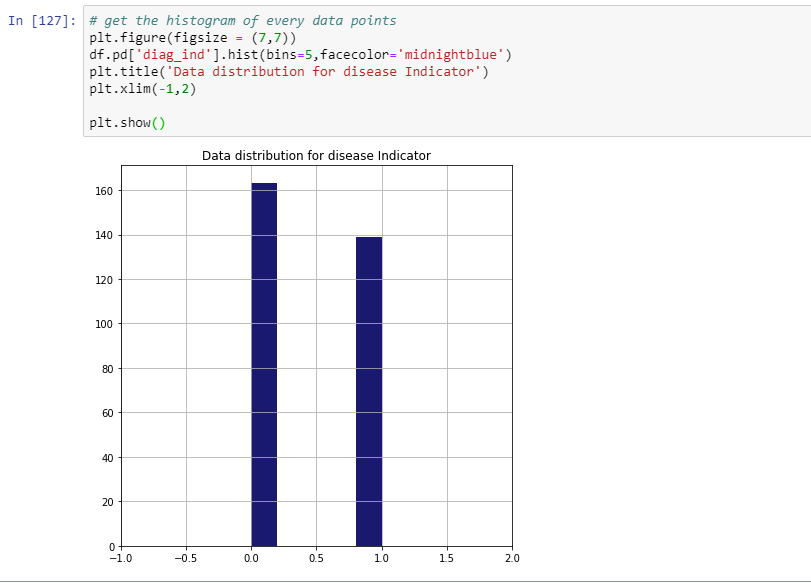
* I can see age distribution is closely resembling of Gaussian distribution.
* I can also find age, blood\_pressure, max\_heart\_rate ,serum\_cholestoral: ,ST\_depression: contains continuous values , while others seems to have categorical values.
* ST\_depression is mostly skewed towards ST\_depression = 0

**5 . Target variable distribution :** Target variable distribution shows that we have equally distributed target data, so we do not need to balance the target field.

Because this is binomial classification , we first create a derived target field such that

# if "diagnosis" == 0, member does not have disease A - we put 0

# if "diagnosis" >= 1, member possess the disease A - we put 1



## **2.4 Verify the Data Quality**

Most of the common errors occur while preparing datasets. This may include spelling mistakes, or data not entered for some records.

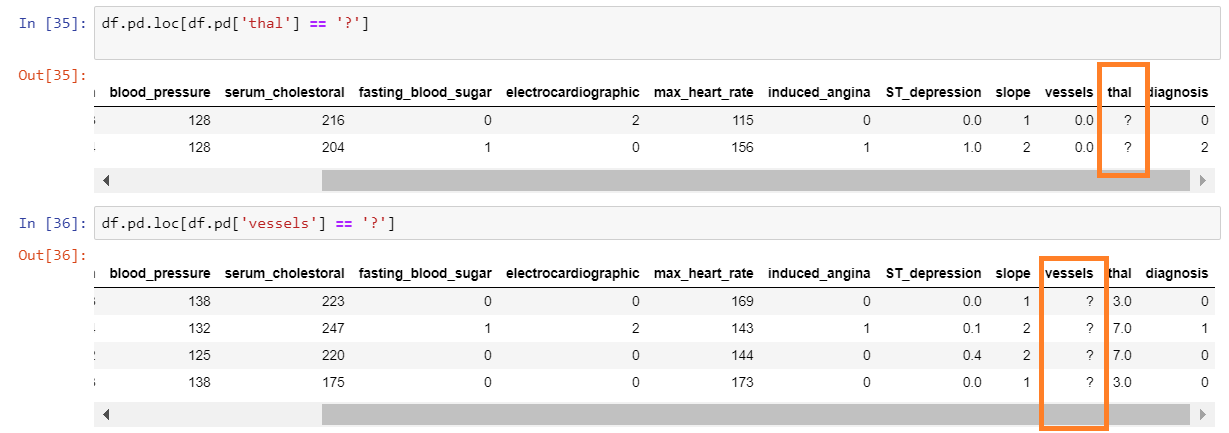
There is a probability of finding many **typographical errors** when dealing with a large dataset. These errors are nothing, but mistakes made during typing, and can occur during data preparation stage. Considering my dataset, the typographical errors can occur only in the String Data. After scanning through my dataset manually, I didn't find any typographical errors, because my data does not contain any string values.

## Another type of error that can occur **is measurement error**, which is the difference between a measured value of a quantity and its true value. It is difficult to expect any measurement errors expected as the data is collected from the authentic providers / Hospitals, which can be regarded as a plausible source of information.

**Missing Values**

In “Describe the Data” stage , I found Most of the datatypes are integer and double , however few of them are string datatype, although they contain numeric values. **That shows , there could be some missing or NA or default string type values.** We will explore them and decide to impute.

I find , that ‘thal’ and ‘vessels’ attributes has some missing values, they are denoted by ‘?’ sign (highlighted in RED)



To clean the missing values, we can make Impute these missing values with mean, median or any other aggregation of the attribute. **I will discuss about cleaning of missing values in the later parts of the report.**

In addition to this , I do not see any **NULL values** for the attributes.



**Flat File Data Quality** – Delimitators are consistent throughout the file , and all the values are properly imported into AWS, Spark environment .

# **3. Data Preparation**

## **3.1 Select the data**

Many of the patient decisions about which data to select have already been made in earlier phases of the data mining process

**Selecting Patients -** The initial study will be limited to the (approximately) 300 patients who have visited to providers supported by our health insurance company ONLY . These patients are chosen random from the pool of patient records .

**Selecting Attributes –** The database contains many sensitive information of patient , such as , PHI (Patient health indicators), SSN (Social security number), Patient ID , name etc . These attributes are removed during the generation of flat file from patient database. These attributes will also not have any impact on our data mining objective.

* There were more data sources available such as processed.Hungarian.data and processed.Switzerland.data files , **However 55% of data in these sources have missing values. Therefore , they are not fit for study**. There were also some missing features in these data sources.
* Column **RecordID** was created manually for this data source, I created these to identify unique instances and utilized for aggregation during data exploration part.

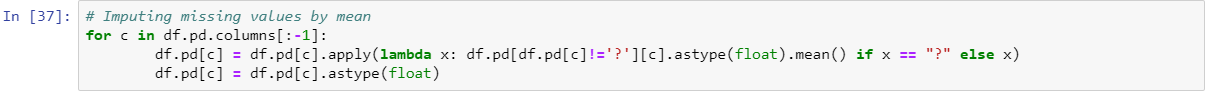
**3.2 Clean the data**

**Missing data -**  As discussed in the previous section, we will perform data cleaning, essentially, imputing missing data.

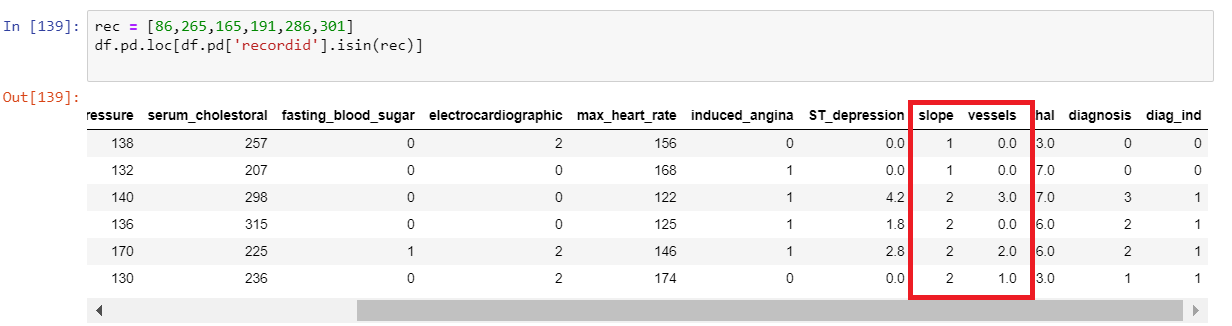
Here are some techniques for missing values imputation

* **Replace with mean**: Calculates the column mean and uses the mean as the replacement value for each missing value in the column. **I will use this technique, because data distribution for these attributes are skewed towards mean, so it is better to replace them by mean values , as number of such missing values are also not large (only 6)**
* **Remove entire row**: Completely removes any row in the dataset that has one or more missing values. This is useful if the missing value can be considered randomly missing**. I will not use this option , because number of missing values are less , and just present in 2 attributes. In addition , I already have less , records.**
* **Replace using Probabilistic PCA**: Replaces the missing values by using a linear model that analyses the correlations between the columns and estimates a low-dimensional approximation of the data, from which the full data is reconstructed**. I will not use this in my dataset, because it is complex to implement in spark and since number of missing values are very small, it is not much ROI.**

Perform missing value imputation.



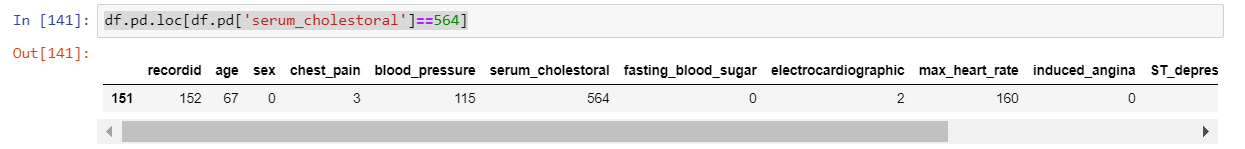
Result after the missing value imputation



**Noise Reduction**

We already found noise in “serum\_cholestrol” attributes in data exploration , we will replace this rows by mean of this attribute.

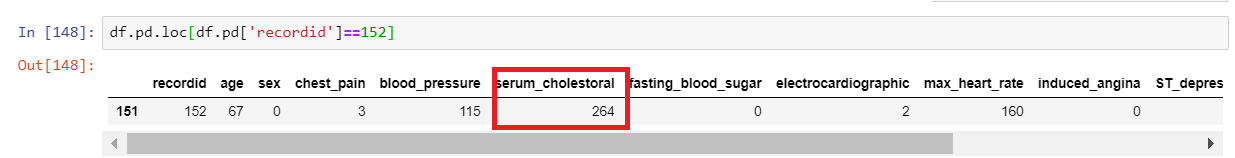
We found in data exploration that **[serum\_cholestrol] = max([serum\_cholestrol]) was the outlier**

****

We will replace this value by **mean([serum\_cholestrol]), which is 264** , we already calculated this mean in attribute statistics in earlier section.

****

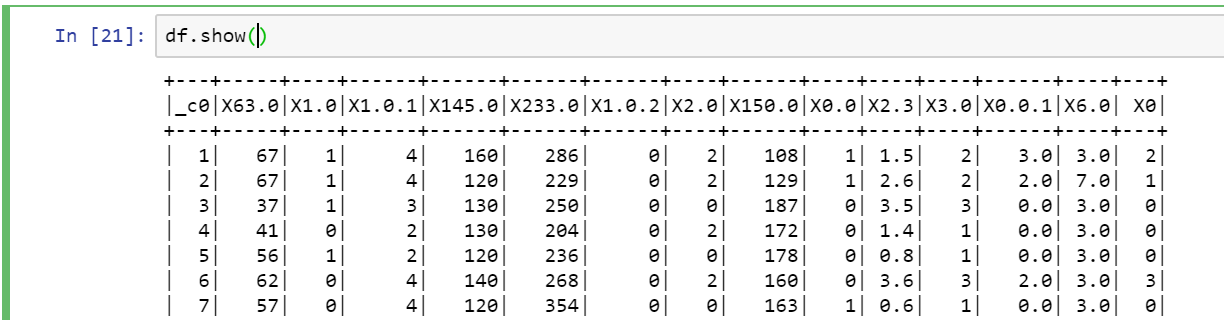
Result after imputing outlier

****

**3.3 Construct the data**

**Naming Features :**

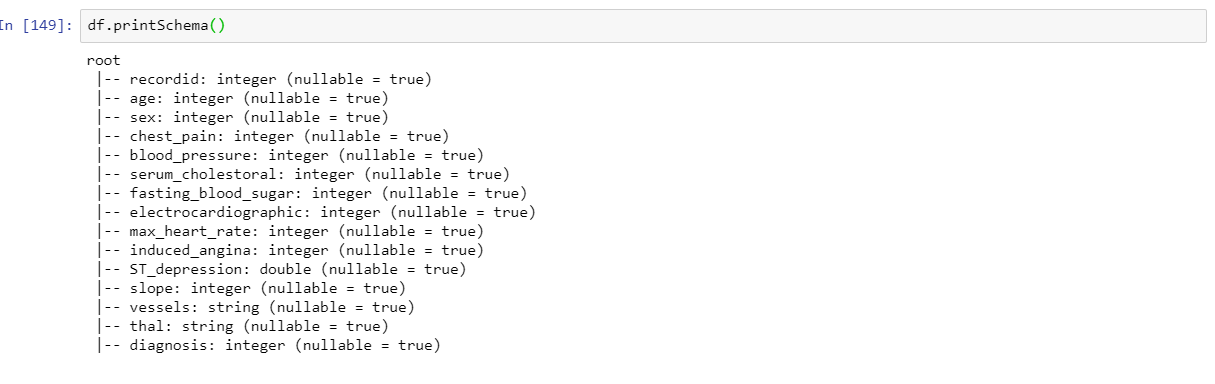
The initial data set do not have feature name defined, Therefore , it is very hard to analyse the data without feature name.



Now , I will give these each attribute a name , based on the data dictionary.



Output after the feature naming



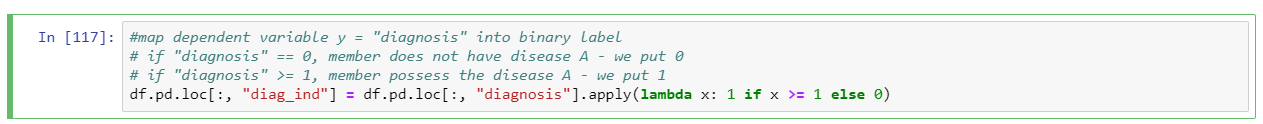
**Derived Attribute :** As we have seen that originally our data contains four different values for **“diagnosis” attribute** , which is our target variable as well. These values are from 0 to 4 depending on angiographic disease status , which are calculated based on diameter narrowing of the heart. **I will name new column as “diag\_ind”**

For this experiment , we have converted these values into 0 or 1 , such as ,

diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing

-- Value 1: > 50% diameter narrowing



**3.4 Integrate the data source**

For now we have taken only one instance (Cleveland) of database, we will add more instance of same database , which contains data from other countries

**3.5 Format the data**

**Change Data types :**

# When data contained missing values , they were denoted by ‘?’. Since ‘?’are considered as string values , the attributes “thal” and “vessels” where shown as **String** datatypes.

After we imputed these missing values and converted pandas dataframe into spark dataframe, they are the attributes “thal” and “vessels” are converted back to double datatype.

# 

**One-Hot encoding :** One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. The categorical value represents the numerical value of the entry in the dataset.

**Since my dataset already has categorical values set as numerical values , I do not need to perform One-hot encoding.**

# **4. Project Data**

* 1. **Reduce the data**

Data Reduction can be performed by selecting relevant features. This is called Feature Selection.

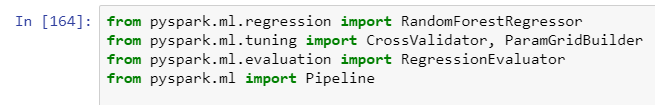
**Feature selection** is the process of selecting a subset of relevant, useful features to use in building an analytical model. Feature selection helps narrow the field of data to the most valuable inputs. Narrowing the field of data helps reduce noise and improve training performance. One way of feature selection is **Feature Ranking.**

**Feature ranking** resembles to some extent to feature selection, in the sense that by ordering features from the most influential (which explains the most variability in the model) to the least influential, one can chose to discard (reduce, eliminate) the latter without impacting too much the final result.

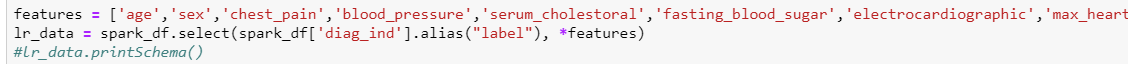
Once again, when we eliminate (reduce, discard) features from your dataset using one or both of these methods, **it is important to have domain knowledge and not rely solely on the statistical and/or mathematical results**.

With that disclaimer in mind, we'll be looking at to rank features using Random Forest Regressor and PySpark.

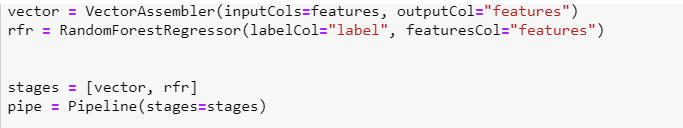
**Step 1** : We have to import some extra libraries



**Step 2** : prepare the data

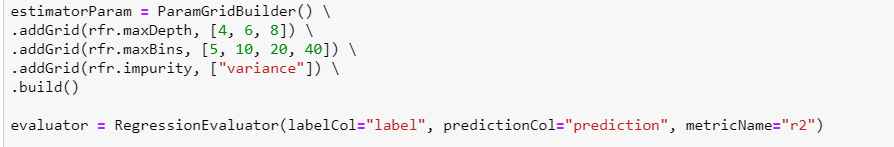


**Step 3** : Next we'll prepare the data pipeline for our random forest regressor



These are all numerical features, so the VectorAssembler are enough to build our pipeline. However, if we had categorical data in our dataset, we would have to use StringIndexer and OneHotEncoder also.

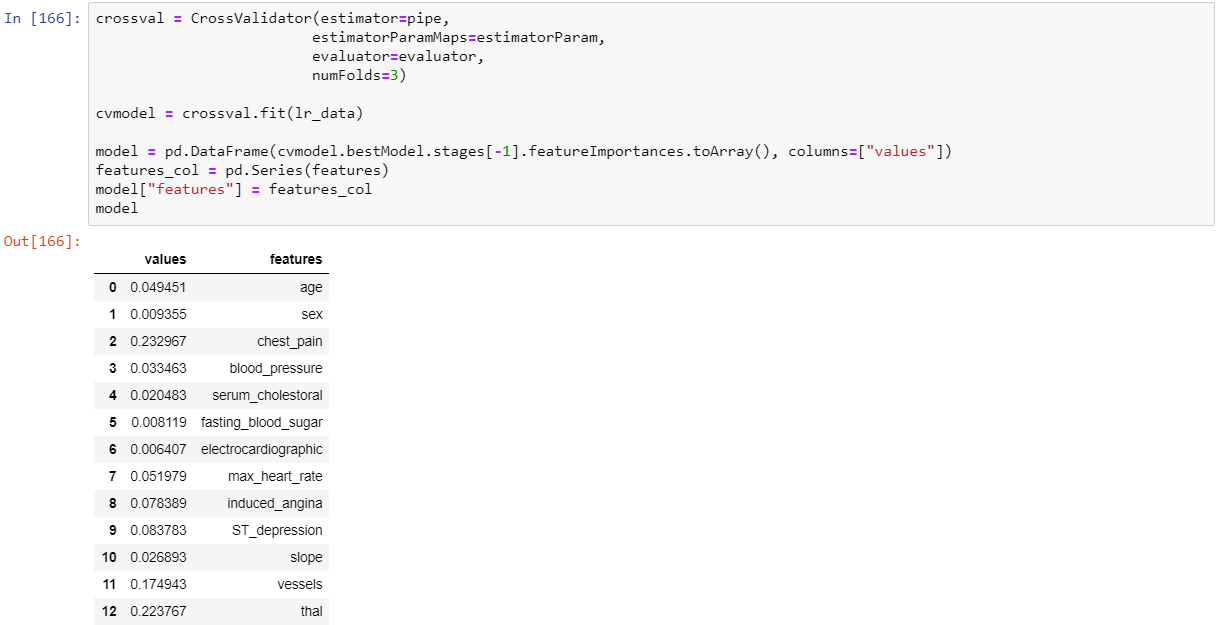
**Step 4** : Now we can build a grid validator and decide on R^2 metric to evaluate our random forest results. Again, domain knowledge is important here when we chose the evaluation metric. Sometimes MAE or mean absolute error may produce better results.



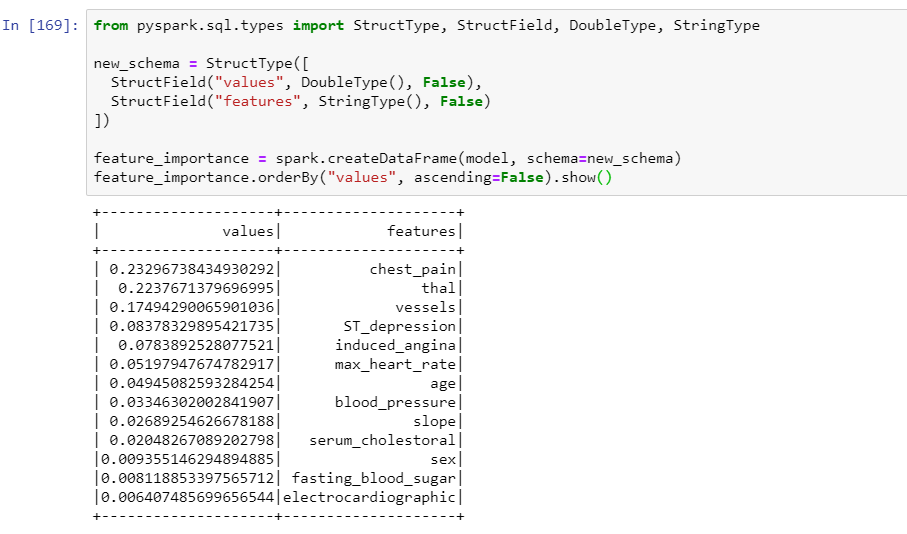
**Step 5 :** The next step is to put everything together and run the cross validation in order to find out which one is the best model out of all the models from the grid.

This part will take a bit of time (probably like 5 min or so) since we have to go through count(maxDepth x maxBins x impurity) worth of model validation.

**Output** : We can get now the best model to show us which features are the most important.



Finally , I display all features , in the descending order of feature ranking.



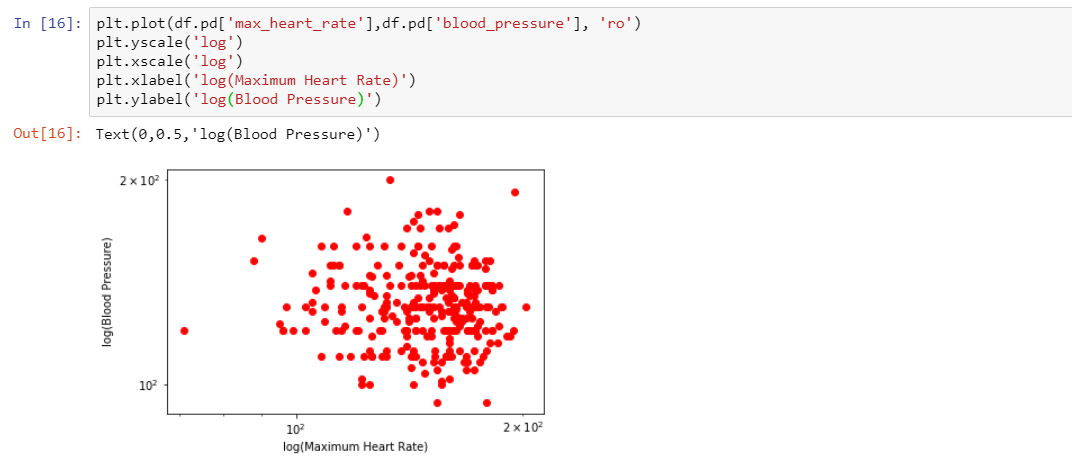
**4.2 Project the data**

Sometimes data projection is required to reduce the dimensionality of data , so that it can be easily interpreted and visualize, such as taking log of one of the attribute and plotting it with another attribute, or taking log of both the plotting attributes , in order to standardize the plotting.

1. Log(max\_heart\_rate) vs Log (blood\_pressure) Plot

Considering max\_heart\_rate and blood\_pressure could be on different scale , I tried to take log values of both the variable and plotted the graph . Although , I do see much change in the distribution , however this method could useful, when my variables are on different scale.

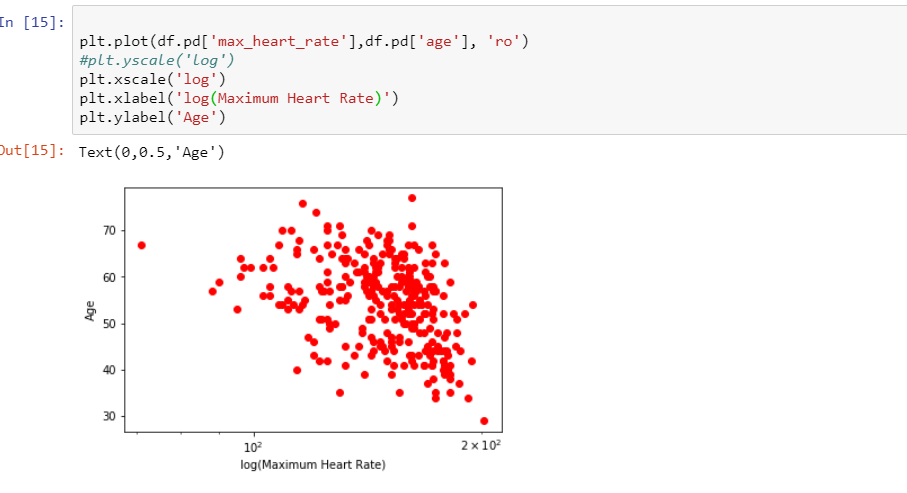
**Most of the data is skewed in top right quadrants, which means , people with higher max\_heart\_rate tend to have higher blood pressure.**



1. age vs Log (max\_heart\_rate) Plot

In this plot , I understand that Age variable Max\_heart\_rate have some scaling , therefore , I plotted the graph of age vs log(max\_heart\_rate)

**I can see, people with higher age tend to have hight max\_heart\_rate**



**Phases – Data Understanding , Data Preparation and Data Transformation have been done in 3 to 4 iteration before moving to next phase.**

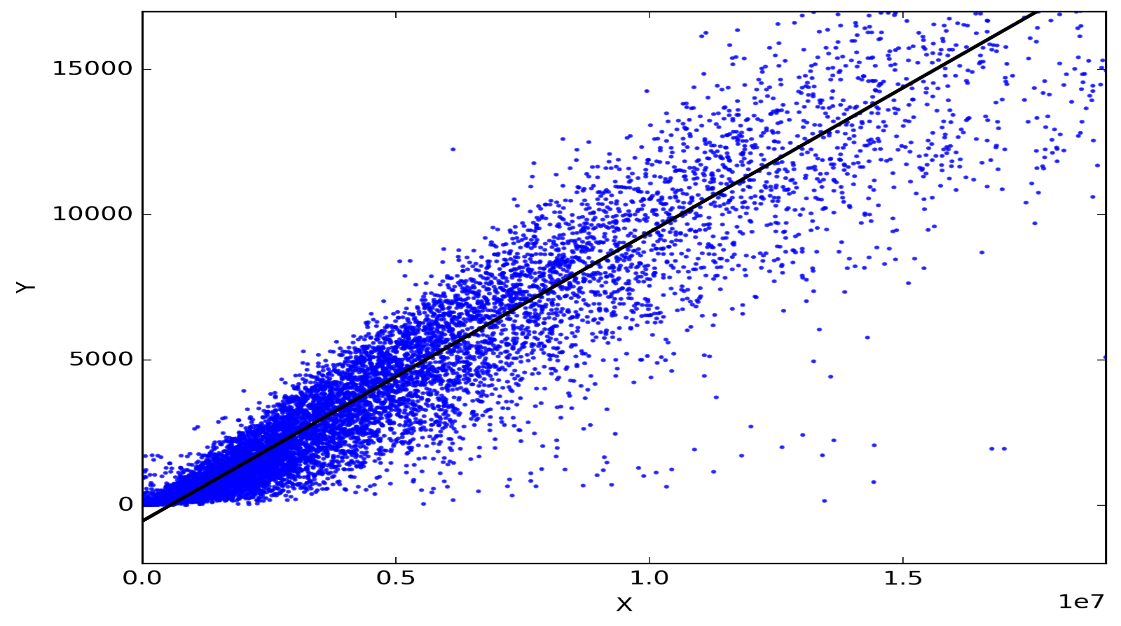
# **5. Data-Mining method(S) Selection**

## 5.1 Data Mining Objective

Data Mining refers to extracting valuable insights from the data. It is a procedure performed after the data has been fully cleaned and transformed. There are multiple data mining techniques that can be applied by a user. Some of the widely used are

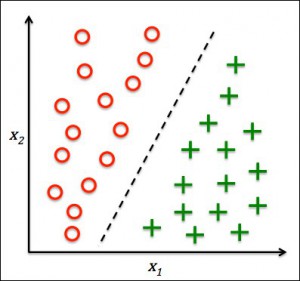
**Regression** - In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

**Note: The following image was taken from** [**www.google.com**](http://www.google.com)**. This image is just for the user’s to give an idea about regression visually**. We see several data points , and the objective of regression problem is to find the best fit curve.



**Classification** - It is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, based on a training set of data containing observations (or instances) whose category membership is known. Examples are assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.).

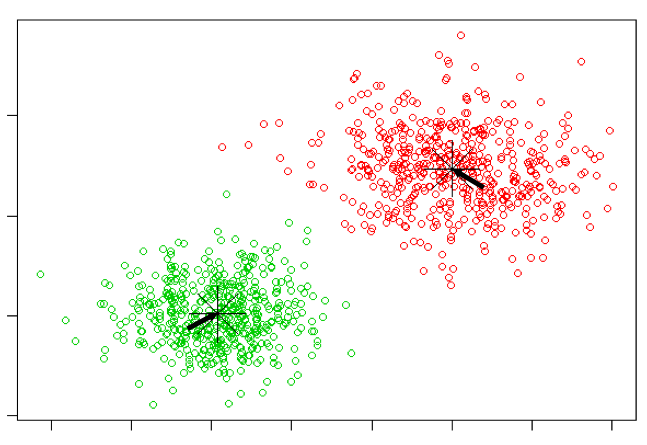
**Note: The following image was taken from** [**www.google.com**](http://www.google.com)**. This image is just for the user’s to give an idea about Classification visually.**



* As per our data mining objective , we want to predict potential heart patients from Target variable ‘Heart\_disease\_ind’ flag , using 14 input variables (health parameters) .
* This problem statement can be categorised as Classification, which is a machine learning method that uses data to determine the category, type, or class of an item or row of data. For example, you can use classification to:
* In addition , since there are only 2 outcomes , 1= patient has heart disease , 0 = patient does not have heart disease , this classification problem can be further classified into **“Binary Classification Problem”.**
* Classification is a supervised machine learning method. It always requires labelled training data

**Clustering** - It is a task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is an unsupervised learning technique, and does not require any target variable to train data on.

**Note: The following image was taken from** [**www.google.com**](http://www.google.com)**. This image is just for the user’s to give an idea about Clustering visually**



## 5.2 Selecting appropriate Data Mining Method

Using appropriate datamining model technique we want to achieve below objectives

* Predict future potential heart patients based on their current health parameters.
* Identify top n health parameters , which are critical indicators of developing heart disease in future.
* Create a datamining model , which can use patient lab data and can consistently predict potential heart patients , so that those can be used for targeted wellness programs.

In addition, different health promotion and training services will be recommended to such subscribers.

* After going through in detail about the working of each technique in the previous section, I came up with following conclusions-
* As per out data mining objective , we want to predict potential heart patients from Target variable ‘diag\_ind’ flag , using 14 input variables (health parameters) , therefore this is a classification problem under Supervised Learning category .
* Also, my data contains target field as binary values, where 0 = Absence of heart disease and 1= presence of heart disease. Hence binary classification is much useful method for out data set.

**Therefore, I choose Binary Classification to be my data mining technique.**

# **6. Data-Mining Algorithm Selection**

Now , after we have decided the type of problem , we would perform steps to choose appropriate model to identify, build and execute correct models.

**6.1 Conduct exploratory analysis and discuss**

* Classification tasks are frequently organized by whether a classification is binary (either A or B) or multiclass (multiple categories that can be predicted by using a single model).Since target variable here has only two outcomes 1 or 0 , therefore this is Binary -Classification problem.
* Given a collection of records (training set )

Each record is by characterized by a tuple (***x***,*y*), where ***x*** is the attribute set and *y* is the class label

***x***: attribute, predictor, independent variable, input

*y*: class, response, dependent variable, output

Task: Learn a model that maps each attribute set ***x*** into one of the predefined class labels *y.*

* Another data mining objective is to identify top health parameters , which are responsible for the heart disease. This problem is more like , finding the attributes/ features , which have high predictive power of classifying a patient for heart disease.

Therefore , as already discussed in the previous analysis, Classification is the most suitable technique to carry out further analysis, **as the objective of my analysis involves prediction, and the target variable in my data is categorical**. In the next step, I will proceed with the selecting algorithms which can help me carry out regression.

**6.2 Selecting Algorithms in a Logical Manner**

Classification problem are solved , using simple to complex machine learning models. Complex model are generally create on top of base classifiers. Some of the base classifiers are

* Base Classifiers
  + Decision Tree based Methods
  + Rule-based Methods
  + Nearest-neighbor
  + Neural Networks
  + Deep Learning
  + Naïve Bayes and Bayesian Belief Networks
  + Support Vector Machines
* Ensemble Classifiers
  + Boosting, Bagging, Random Forests

**Two Class Logistic Regression :**

* Logistic regression is a well-known method in statistics that is used to predict the probability of an outcome, and is especially popular for classification tasks. The algorithm predicts the probability of occurrence of an event by fitting data to a logistic function.
* To train this model, you must provide a dataset that contains a label or class column. Because this module is intended for two-class problems, the label or class column must contain exactly two values.

**Multi Class Decision Jungle :**

[Decision jungles](http://go.microsoft.com/fwlink/?LinkId=403675) are a recent extension to [decision forests](http://go.microsoft.com/fwlink/?LinkId=403677). A decision jungle consists of an ensemble of decision directed acyclic graphs (DAGs).

Decision jungles have the following advantages:

* By allowing tree branches to merge, a decision DAG typically has a lower memory footprint and a better generalization performance than a decision tree, albeit at the cost of a somewhat higher training time.
* Decision jungles are non-parametric models, which can represent non-linear decision boundaries.
* They perform integrated feature selection and classification and are resilient in the presence of noisy features.
* **Therefore , I choose only one type of classification algorithm , Logistic regression , because my data mining objective was not to compare different classification algorithm, rather it was identifying best model parameters and appropriate features , which can help to identify future heart patients.**
* I will be running Logistic Regression , with different model parameters and changing the feature , by reducing feature of least importance in every next iteration then decide which technique to choose to deploy.
* **I will also use Random Forest Regressor , to identify feature importance by ranking each feature based on confidence probability to get top ‘n’ features. to give the best result , without worsening the overfitting problem.**

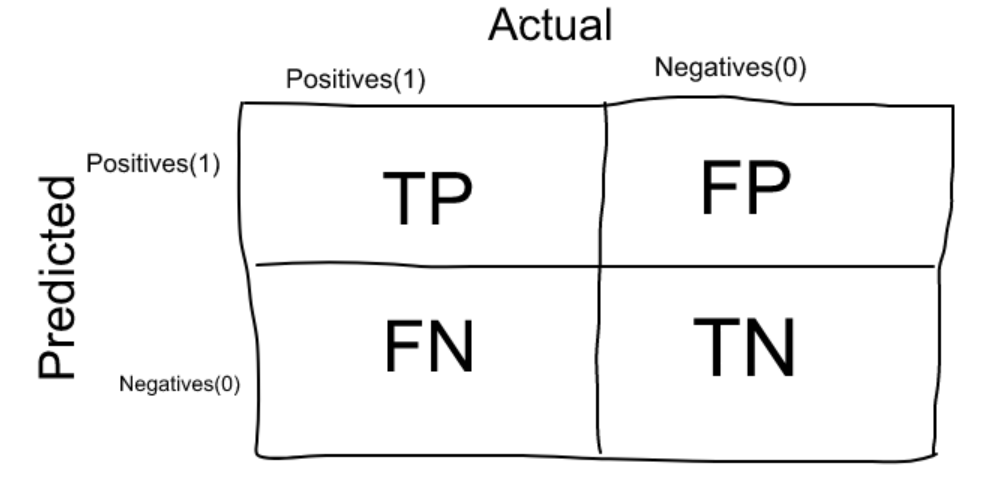
**6.3 Build/Select appropriate model(s) and choose relevant parameter(s)**

* Now we start designing and building the model , Starting with importing the data and then performing data preparation using several data transformation and meta data setup.
* I will also utilize feature ranking , which I derived from feature selection section to optimize my model parameters.

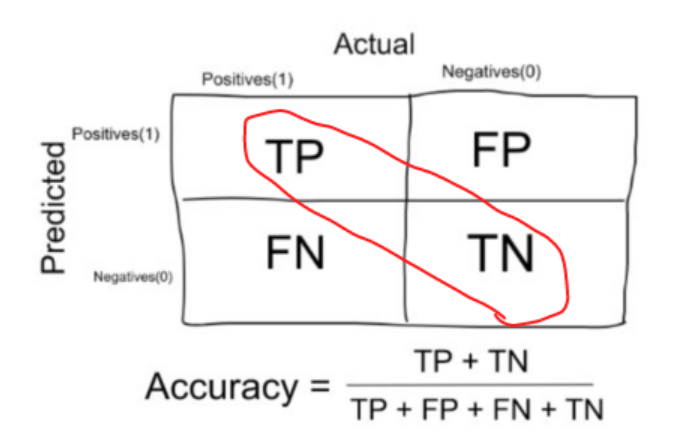
After I perform Data Mining, I will analyse Model Score one by one. Most of the techniques will give the following output. The metrics that you choose to evaluate your machine learning model is very important. Choice of metrics influences how the performance of machine learning algorithms is measured and compared.

**Confusion Matrix:**

The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes.



**Accuracy**: Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.



Accuracy is a good measure when the target variable classes in the data are nearly balanced. **The target variable in our dataset is balanced , Therefore , Accuracy is going to be most important metrics for model evaluation.**

* Once the data is set up , next we split the data using the test design strategy in Step 7. We have created Two-Class Logistic Regression model and Multi Class Decision jungle model for this classification problem.

**Iterations and Parameters**

**Iteration 1 :** Logistic Regression

**Model Algorithm** : Two Class Logistic Regression

**Features** : All Features

**Model Parameter –**

* elastic\_net\_param : L2
* reg\_param : default
* max\_iter : 100
* threshold: default

**Iteration 2**

**Model Algorithm** : Two Class Logistic Regression

**Features** : Top 9 features obtained from Feature selection *['thal','vessels','ST\_depression','induced\_angina','max\_heart\_rate','age','blood\_pressure','slope','serum\_cholestoral']*

**Model Parameter –**

* elastic\_net\_param : L2
* reg\_param : default
* max\_iter : 12.
* threshold: default

**Iteration 3**

**Model Algorithm** : Two Class Logistic Regression

**Features** : All Featurees

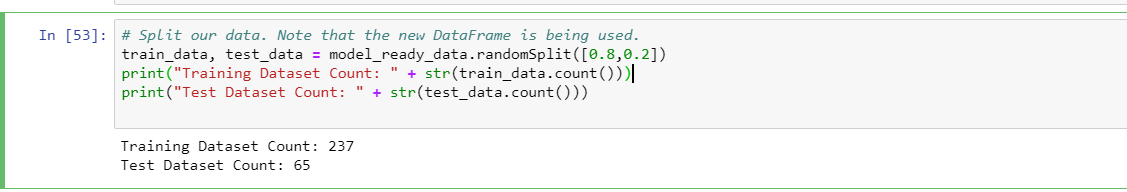
**Model Parameter –**

* elastic\_net\_param : L2
* reg\_param : default
* max\_iter : 50
* threshold: default

# **7. Data Mining**

**7.1 Create and justify test designs**

* In order to test the models and its iterations, we will have to split the data into training and test datasets, using Split data task.
* For this study , 80% data will be used for training and 20% data for test, because machine learning algorithms learn better with more data, this reduces overfitting and improves the accuracy.
* In order to measure the success of model – we would evaluate metrics such as Confusion metrics, Accuracy , AUC curve , Gains.



7.2 Conduct data mining

Now after building the model and setting up the test design , the model is executed. Below are several screenshots of parameter setup and their execution

**Iteration 1** : In this iteration , we will consider all the features and default parameters

* elastic\_net\_param : ElasticNet mixing parameter, in range [0, 1]. For alpha = 0, the penalty is an L2 penalty. For alpha = 1, it is an L1 penalty. Default is L2.
* reg\_param : Regularization parameter (aka lambda)
* max\_iter : The maximum number of iterations to use. Default is 100.
* threshold: in binary classification prediction, in range [0, 1].

****

**Iteration 2 :** In this iteration , we will **consider Top 9 feature**s , which we got from feature selection and will only change **maxIter = 12** parameter, rest all parameters will be used as default parameter. Other parameters are as follows.

* elastic\_net\_param : L2
* reg\_param : default
* max\_iter : 12.
* threshold: default

It is evident , that sometimes **when the iteration increases , problem of overfitting increases**, therefore , I want to try this iteration by reducing maxiter parameter.

****

**Iteration 3 :**

In this iteration , we will consider all the features and maxIter = 50. Although, in the last iteration , I got much better accuracy and least overfitting, this time , I want to see , if I include all the parameters and reduce maximum iteration , does my accuracy increase ?

Model parameters

* elastic\_net\_param : L2
* reg\_param : default
* max\_iter : 50.
* threshold: default

****

**I have executed , above set up , with multuple other iteration , by changing other parameters as well. Since I was not getting any significant differences , therefore I decided to put only those iterations in this document, where we see clear differences in outcome.**

**7.3 Search for patterns**

Now, after the models are executed , the results of the model are can be seen using Evaluation model task

* Below are the results of Two class Logistic Regression model , when ran across 3 iterations. We find results of all 3 iterations in term of 6 metrics
* This model gives best results , in iteration 1 , when are **when we reduce the maxIter from default value , with an accuracy of 89% and highest area under curve**, which refers to the amount of area under the ROC curve (plot between True positive rate and False positive rate)
* **I also observed , that model does not give best result , when I included all the features, instead it gave me best results , when I choose top 9 features, from feature selection ranking. I choose only top 9 features, because feature importance score for top 9 was similar and difference of feature importance score from next 5 features was very high.**
* The average log loss of Iteration 2 is also minimum for this model , measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The goal of our machine learning models is to minimize this value. A perfect model would have a log loss of 0.
* In iteration 3 , the accuracy reduced drastically , It looks like , **there is problem of overfitting**. I will check this in next section .
* Model results make sense on the first glance , however Flase positives and False negatives needs further exploration.
* **Storage :** Model results can also be stored into data warehouse in form of Score labes, with their probabilities . These results can be used wellness programs’s classification engine to identify possible patients , who can be enrolled in wellness program , in order to reduce the risk of heart disease.

Overall , Logistic regression with with default parameter except reduced maxIter and with top 9 features gives the best accuracy among all the iterations.

# **8. Interpretation**

**8.1 Study and discuss the mined patterns**

At this stage, we formalize our assessment of whether or not the project results meet the business success criteria. This step requires a clear understanding of the stated business goals, so be sure to include key decision makers in the project assessment

**Ranking the Models :** Because several of the initial models seemed to make business sense, ranking within that group was based on statistical criteria, ease of interpretation, and diversity. Therefore model has been ranked according to the data mining objective , and in accordance to business understanding.

**New Questions.** : The most important question to come out of the study is, How does external factors such as smoking / alcohol consumption, can affect other health parameters , which are not taken into consideration during this data mining project and can cause heart disease.

Therefore , I choose only one type of classification algorithm , Logistic regression , because my data mining objective was not to compare different classification algorithm, rather it was identifying best model parameters and appropriate features , which can help to identify future heart patients.

We identified , best features from feature ranking algorithm of Random forest Regressor and model parameters , when we iterated different model of logistic regression , calibrating parameters in each iteration.

**8.2 Visualize the data, results, models, and patterns**

**Two Class Logistic Regression model evaluation :**

Logistic regression models can be evaluated using ROC curves , and Precision/Recall plot. These plots are produced after every run in each iteration 1 to 3 .

* The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal).
* For every possible cut-off point or criterion value you select to discriminate between the two populations, there will be some cases with the disease correctly classified as positive (TP = True Positive fraction), but some cases with the disease will be classified negative (FN = False Negative fraction). On the other hand, some cases without the disease will be correctly classified as negative (TN = True Negative fraction), but some cases without the disease will be classified as positive (FP = False Positive fraction).
* A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.

**Iteration 1 :**

* In this iteration we see 92% of the area is covered by this curve , which looks to be very significant, however **I see some overfitting , because the accuracy of model on test data is just 84%. We will compare this curve , with other iterations**.

|  |  |
| --- | --- |
|  |  |

Similarly , The perfect test will have a PRC that passes through the upper right corner (corresponding to 100 % precision and 100 % recall). Generally you can say that the closer a PRC is to the upper right corner, the better the test is.

We will compare this curve with other iteration , and compare , which iteration has this curve more nearer to upper right corner.

**Iteration 2**

As, discussed above , in this iteration , when we reduced the number of iteration and we also did not included all the features , then both ROC and PRC curves outperform than iteration 1.

|  |  |
| --- | --- |
|  |  |

The precision vs Recall curve is also more inclined towards upper right corner in this iteration than iteration 1, which signifies better model performance than iteration 1 , to capture TP, FP, FN.

**The accuracy of this model on test data is 89% and AUC on model summary is 90% . Therefore , we can conclude that the problem of overfitting, which we encountered in iteration 1 is now reduced significantly. This model performance is better than iteration 1.**

**Iteration 3 :**

In this iteration , we took median of maxIter parameter as compared to iteration 1 and 2 , and also included all the features in the model.

**In this case we find best AUC and Precision-recall curve among all the iteration. However one thing to note is that Accuracy on the test data is just 72%, lowest in all the iteration. Therefore , I concluded that problem of overfitting got worse , when we include all features and reduce maxIter parameter.**

|  |  |
| --- | --- |
|  |  |

**8.3 Interpret the results, models, and patterns**

In this section, we will combine the results of above 2 section and list insights.

**Two Class Logistic Regression model evaluation :**

* **ROC Curve :** Usually when we see two models compared in a graph as the one above, the lines only intersect in the corners. The line which comes closest to the upper-left corner provides the best predictions. For every threshold value it scores better than the other model for both false positives and false negatives. The further a curve is to the middle, the worse is its performance.

For this data set, Iteration 2 curve is most nearer to the upper left corner , hence it is closest to best fit model.

For analytic purposes, we are looking for curves that are closer to the top left corner.  We see that the more "efficient" model is iteration 1 of this model.

* **Precision / Recall plot -** To understand how this graph relates to false positives and false negatives, you need to remember that Precision and Recall are calculated as follows:

**Recall**= True Positives / (True Positives + False Negatives)

**Precision**= True Positives / (True Positives + False Positives)

In this curve, the “sweet spot” for the ideal model is in the upper right corner. In iteration 1 , the curve is much more near to best fit as compared to iteration 2 and 3 , which shows better fit for the model

**There were many interesting outcomes from the above sections.**

* It is clearly seen , Iteration 2 has lowest False negative and False positive , which is lowest among the iteration 1,2,3. Therefore , it increases the overall accuracy of this model to 90%, better than any other model for this study.
* Features : **thal', 'vessels', 'ST\_depression', 'induced\_angina', 'max\_heart\_rate', 'age', 'blood\_pressure', 'slope', 'serum\_cholestoral'** are the most important feature for predicting if a person has a heart disease.
* Although , blood pressure , cholesterol are highly correlated to the heart disease in theory , however they have very least feature importance for predicting a heart disease. That’s sounds very interesting and need more domain knowledge to get further insights.
* **Overfitting ,** which means that models performs better on the training data , but not that good on the test or validation data, showed significant shift , when I actually changed the model parameter and features.
* For example - Accuracy on the test data is just 72% in iteration 3, lowest in all the iteration, however it has highest AUC Therefore , I concluded that problem of overfitting got worse , when I include all features and reduce maxIter parameter.
* **One interesting thing was observed , that Right feature selection has high impact on the accuracy and overfitting of the model , as compared to model parameter**. By changing the model parameters , did not gave us much difference in the results , but when I selected features from high importance to low importance , significant differences were seen in accuracy and overfitting.
* **False Negatives :** This metric is very important for my data mining objective , because we do not want to mis those people , who actually have heart disease and my model does not predict that. It is always worth to go for medical testing to confirm presence of disease and its severity, However , if the model misclassify such people , they would miss regular check up and a minor problem can even get worse.

**8.4 Assess and evaluate results, models, and patterns**

**Conclusion / Result** :

* In this document the problem of constraining and summarizing different algorithms of data mining used in the field of medical heart disease prediction are discussed. The focus is on using different algorithms and combinations of several features for intelligent and effective heart attack prediction using data mining.
* For predicting heart attack, significantly 14 attributes are listed and with basic data mining technique other approaches e.g. Clustering and Association Rules, soft computing approaches etc. can also be incorporated.
* The outcome of predictive data mining technique on the same dataset reveals that Logistic regression with different maxIter parameter and other default parameter is having similar accuracy but, when we reduce maxIter parameter from default 100, then accuracy improves marginally.
* The second conclusion is that, appropriate feature selection has high impact on accuracy and overfitting. Including all the features with Logistic regression classification , does not gives best result and make overfitting problem worse . There are only 9 features **thal', 'vessels', 'ST\_depression', 'induced\_angina', 'max\_heart\_rate', 'age', 'blood\_pressure', 'slope', 'serum\_cholestoral' ,** when used gives best accuracy and least overfit model.
* So at the end of this exercise we could comfortably identify few best suited model and finally choose the best model after evaluation for our data mining objective. Therefore to corelate our findings with our data mining objective , we can say -
* Given 9 health parameters that, **thal', 'vessels', 'ST\_depression', 'induced\_angina', 'max\_heart\_rate', 'age', 'blood\_pressure', 'slope', 'serum\_cholestoral'**. we can predict potential future heart patient by 90% accuracy and we can target these insurance subscribers with better health insurance plan , recommend wellness programs and suggest ways for course-correction in order to stay healthy.

**Scope of Improvement**

* We have not taken any external factors such as smoking / alcohol consumption into consideration in our data set that can affect other health parameters , which are not taken into consideration during this data mining project and can cause heart disease.
* The dataset volume is not large , however , we know that LSVM performs better with large datasets, so we can take larger dataset and test our model to get more confidence in our approach using LSVM
* This study was limited to only 9 input health parameter, this can be scaled further to add some more input variable and study the impact on model performance , or choose another best performing model.
* We can also build blending/stacking models. Blending or stacking refers to the method(s) where models take the predicted values of other models as predictors. These are called ensembles and sometimes giver better accuracy than individual models. Random forest , which is an ensemble method , can be used to perform more iteration in order to gain better performance.

**8.5 Iterate prior steps (1 – 7) as required**

All the steps are performed in multiple iterations whenever the results were not quite, optimal, we consider another round of modelling.