****

|  |  |
| --- | --- |
| **Course Number** | **CBD 2214 1** |
| **Course Title** | **BIG DATA FUNDAMENTALS – DATA STORAGE NETWORKING** |
| **Semester/Year** | **Winter 2019** |

|  |  |
| --- | --- |
| **Instructor** | **Parisa Naraei, PhD** |

|  |
| --- |
| **MACHINE LEARNING A-Z REPORT** |

|  |  |
| --- | --- |
| **Due Date** | **March 28/2019** |

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Student ID** | **Signature\*** |
| **SHREE MAHALAKSHMI** | **C0748070** | **SM** |
| **SAI BALAKRISHNA** | **C0754209** | **SB** |
| **ANJU THANKACHAN** | **C0746314** | **AT** |
| **RAVINDER KAUR** | **C0754211** | **RK** |

Table of Contents

[**Cover page** 0](#_Toc4679114)

[Table of Figures 1](#_Toc4679115)

[1.Introduction: 2](#_Toc4679116)

[2.Dataset: 2](#_Toc4679117)

[3.Data Preparation: 3](#_Toc4679118)

[3.1. Obtaining Data: 3](#_Toc4679119)

[3.2. Exploratory Data Analysis: 3](#_Toc4679120)

[3.3 Data Preprocessing 4](#_Toc4679121)

[4. Feature Engineering: 5](#_Toc4679122)

[5.Modeling: 5](#_Toc4679123)

[6.Performance Estimation: 6](#_Toc4679124)

[References 7](#_Toc4679125)

# Table of Figures

[Figure 1 - EDA plotting before cleansing 3](#_Toc4670822)

[Figure 2 - Boxplot of trestbps before outlier removal 4](#_Toc4670823)

[Figure 3 - Boxplot of trestbps after outlier removal 4](#_Toc4670824)

# 1.Introduction:

Through this project, we have attempted to predict the diagnosis of heart disease using the dataset provided. We employed machine learning capabilities to achieve the goal and analyzed the performance of the model.

# 2.Dataset:

For this process, we have used the Cleveland dataset obtained from the heart disease database. This dataset consists of 14 attributes which are considerably linked to the heart disease. Following are the details of attributes available in the dataset:

|  |  |  |
| --- | --- | --- |
| No. | ATTRIBUTES | DETAILS |
| 1 | Age | Age in years |
| 2 | Sex | sex (1 = male; 0 = female) |
| 3 | Chest pain | chest pain type  -- Value 1: typical angina  -- Value 2: atypical angina  -- Value 3: non-anginal pain  -- Value 4: asymptomatic |
| 4 | Trestbps | resting blood pressure |
| 5 | Chol | serum cholestoral in mg/dl |
| 6 | Fbs | (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) |
| 7 | restecg | resting electrocardiographic results  -- Value 0: normal  -- Value 1: having ST-T wave abnormality (T wave inversions and ST  elevation or depression of > 0.05 mV)  -- Value 2: showing probable or definite left ventricular hypertrophy  by Estes' criteria |
| 8 | Thalach | maximum heart rate achieved |
| 9 | Exang | exercise induced angina (1 = yes; 0 = no) |
| 10 | Oldpeak | ST depression induced by exercise relative to rest |
| 11 | Slope | the slope of the peak exercise ST segment  -- Value 1: upsloping  -- Value 2: flat  -- Value 3: downsloping |
| 12 | ca | number of major vessels (0-3) colored by fluoroscopy |
| No. | ATTRIBUTES | DETAILS |
| 13 | Thal | 3 = normal; 6 = fixed defect; 7 = reversable defect |
| 14 | Num | diagnosis of heart disease (angiographic disease status)  -- Value 0: < 50% diameter narrowing  -- Value 1: > 50% diameter narrowing |

# 3.Data Preparation:

## 3.1. Obtaining Data:

We have imported the dataset using Python Pandas packages which transform our dataset into Pandas Dataframe to perform the required operations. Column names were passed as a parameter while importing the dataset since those details were missing in the actual dataset.

## 3.2. Exploratory Data Analysis:

Exploratory Data Analysis is performed to obtain knowledge about the dataset to identify the pattern and gather insights from it. By plotting the graph with the data provided, we can comprehend the relation between each attribute and the dependent variable (num)

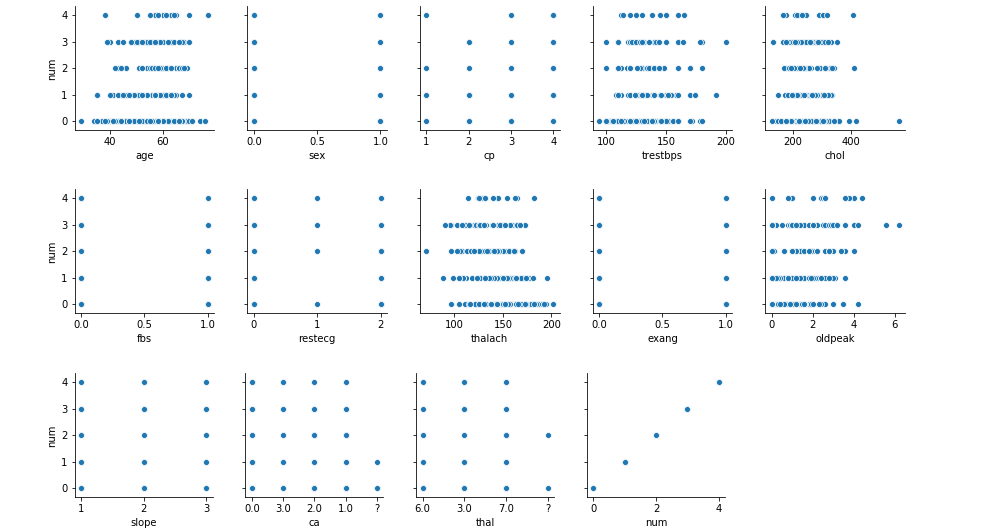


Figure 1 EDA plotting before cleansing

## 3.3 Data Preprocessing

In data preprocessing we have detected and handled missing data and outliers.

* Missing data were handled by replacing with the mean value of the column.
* Outliers in the dataset have an impact on the result. To obtain maximum performance we need to handle the outliers. We have detected the outliers in columns 'cp','trestbps','chol','fbs','thalach','old peak' and 'ca' by using the boxplot and those outliers were removed.

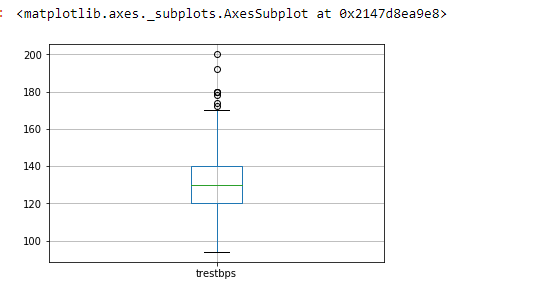


Figure 2 - Boxplot of trestbps before outlier removal

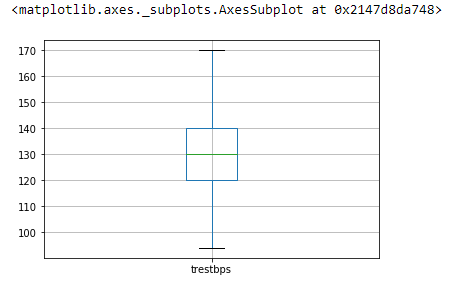


Figure 3 - Boxplot of trestbps after outlier removal

* There were no redundancy in the dataset.
* The ‘num’ column refers to the presence of heart disease in the patient which has five values ranging from 0 to 4. Here 0 indicates the absence of the disease whereas 1 to 4 represents the presence. Since the values 1 to 4 falls under the same category, we have grouped 1 to 4 as 1 for easier processing.

# 4. Feature Engineering:

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data (Bhandare, 2017). It includes

* Feature Construction
* Feature transformation
* Dimension reduction

Since our dataset was already processed to some extent, we are not performing feature engineering steps. All 14 attributes contribute to predicting the target.

# 5.Modeling:

In data modelling, we split the data into train data and test data.

• Training data is the data which is used to fit into the classifier and analyzed.

• Test data is used to test a model and it is randomly split into subset and excluded from training process. It is used to determine the performance and accuracy of the classifier.

We have split our dataset into 80% of train data and 20% of test data.

Since we have the independent variable in dataset and our goal is to predict the classes we have used "Logistic Regression"

The accuracy obtained using this classifier is 87%

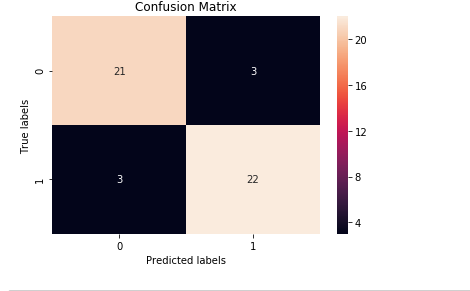
# 6.Performance Estimation:

Cross Validation:

We have used K-Fold for cross validation since we train only on few observations the classifier will perform well and leads to over estimation. Here we have split data into 10 folds. The mean accuracy using cross validation is 82%

Confusion matrix:

Confusion marix to determine all combinations of pricted and actual values. We have measured recall,precision an accuracy.



# References

Bhandare, S. G. (2017). Bhandare, S., Goldberg, D., & Dowell, R. (2017). Discriminating between HuR and TTP binding sites using the k-spectrum kernel method. PLoS One, 12(3), e0174052.