ASSIGNMENT-2

Part B: Assignments based on R and Python

Aim:

Perform the following operations using R/Python on the Air quality and Heart Diseases data sets

1) Data cleaning
2) Data integration
3) Data transformation
4) Error correcting
5) Data model building

Introduction.

The Air Quality Dataset is available in R Studio and the Operation are performed on this data.

Aim of analysis

In the following document, 4 different machine learning algorithms to predict heart disease (angiographic disease status) are compared. For some algorithms, model parameters are tuned and the best model selected. The best results measured by AUC and accuracy are obtained from a logistic regression model (AUC 0.92, Accuracy 0.87), followed by Gradient Boosting Machines. From a set of 14 variables, the most important to predict heart failure are whether or not there is a reversable defect in Thalassemia followed by whether or not there is an occurrence of asymptomatic chest pain.

Dataset:

Nicely prepared heart disease data are available at UCI. The document mentions that previous work resulted in an accuracy of 74-77% for the preciction of heart disease using the cleveland data.

Variable name	Short desciption	Variable name	Short description
age	Age of patient	thalach	maximum heart rate achieved
sex	Sex, 1 for male	exang	exercise induced angina (1 yes)
cp	chest pain	oldpeak	ST depression induc. ex.
trestbps	resting blood pressure	slope	slope of peak exercise ST
chol	serum cholesterol	ca	number of major vessel
fbs	fasting blood sugar larger 120mg/dl (1 true)	thal	no explanation provided, but probably thalassemia (3 normal; 6 fixed defect; 7 reversable defect)
restecg	resting electroc. result (1 anomality)	num	diagnosis of heart disease (angiographic disease status)

The variable we want to predict is **num** with Value 0: < 50% diameter narrowing and Value 1: > 50% diameter narrowing. We assume that every value with 0 means heart is okay, and 1,2,3,4 means heart disease.

From the possible values the variables can take, it is evident that the following need to be dummified because the distances in the values is random: cp,thal, restecg, slope.

Operations Performed on Air Quality Dataset

Read the Dataset

```
> airquality
    Ozone Solar.R Wind Temp Month Day
1      41     190  7.4   67     5   1
2      36     118  8.0   72     5   2
3      12     149 12.6   74     5   3
4      18     313 11.5   62     5   4
5      NA      NA 14.3   56     5   5
6      28      NA 14.9   66     5   6
7      23     299  8.6   65     5   7
8      19      99 13.8   59     5   8
9       8      19 20.1   61     5   9
10     NA     194  8.6   69     5  10
11      7      NA  6.9   74     5  11
12     16     256  9.7   69     5  12
13     11     290  9.2   66     5  13
14     14     274 10.9   68     5  14
15     18      65 13.2   58     5  15
```

View the Summary of Data

```
> summary(airquality)
     Ozone           Solar.R          Wind             Temp      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
 NA's   :37       NA's   :7                                       
     Month            Day       
 Min.   :5.000   Min.   : 1.0  
 1st Qu.:6.000   1st Qu.: 8.0  
 Median :7.000   Median :16.0  
 Mean   :6.993   Mean   :15.8  
 3rd Qu.:8.000   3rd Qu.:23.0  
 Max.   :9.000   Max.   :31.0  
```

1) Data cleaning (Students have to Insert the screenshot/output of every options operations Performed by them.)

Conclusion: Thus we have learnt various operations of 1) Data cleaning 2) Data integration 3) Data transformation 4) Error correcting 5) Data model building) with **R Language in RStudio**.