**Classification of Patients with Abnormal Blood Pressure**

**Background**

**Hypertension** with aging is a major medical concern even in this ultramodern era of advanced therapies. Preliminary testing is a key element in analyzing the procedures of mild abnormality of blood pressure for a substantial period of time, but presently has a limited value in the prediction of progression to hypertension. Although **hypotension**, low blood pressure in common parlance, is less common among the ones with abnormal blood pressure, but the adverse effect it has on their health is equivalent to that of hypertension.

Although recent studies have hypothesized that, keeping other factors constant, hemoglobin level is positively associated with blood pressure in a large cohort of healthy individuals, but genetic causes are also prominent in some of individuals. However, there are numerous other factors which determine whether an individual is likely to develop this abnormality.

Data scientists lay out the hypothesis that newer statistical classification methods derived from data mining and machine learning methods are capable of reducing the prediction error manifolds and help cardiologists to conduct a two-tailed preliminary prediction of abnormality of blood pressure in an individual.

**Objective**

Employing statistical techniques, conduct a preliminary prognosis of Hypertension/hypotension, based on the level of **hemoglobin** and **genetic** **history** of the individual.  
*[Please refer to the Data Dictionary (on the next page) to know more about these variables]*

**Deliverables**

1. Lay out an **approach plan**, consisting of:
   1. Your **understanding of data**, based on a preliminary exploratory analysis
   2. **Different** traditional as well as state-of-the-art statistical **techniques**, which you are going to use to come up with different models to meet the objective
2. Contrast the **pros** **and cons** **of** applying each **technique** on this problem
3. **Build a model** using the most promising technique on the dataset.

**Model Validation** is supposed to be done on the test dataset *(to be given to you during your case-study presentation)*

1. What would be your **approach**, **if** there were **other variables also** in the data:

*Smoking, obesity (BMI), Lack of physical activity, salt content in the diet, alcohol consumption per day, Level of Stress, Age, Sex, Pregnancy, Chronic kidney disease* and *Adrenal & thyroid disorders*.

*[Please refer to the Data Dictionary (on the next page) to know more about these variables]*

Share the software code used to execute each technique or operation.

The approach plan, model outputs, model diagnostics, software codes etc. are supposed to be shared in an Excel file.

Support your deliverables with exhibits and slabs, wherever required.

***Note: The data are hypothetical.***

**Data Dictionary**

**Total** Number of Patients (N) = 2000

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Position** | **Variable Label** | **Value Labels** | **Measurement Level** | **Role** |
| **Patient\_Number** | 1 | Patient Number | Not Applicable | Ratio | None |
| **Blood\_Pressure\_Abnormality** | 2 | Blood Pressure Abnormality | 0 = Normal | Nominal | Target |
| 1 = Abnormal |
| **Level\_of\_Hemoglobin** | 3 | Level of Hemoglobin (g/dl) | Not Applicable | Ratio | Input |
| **Genetic\_Pedigree\_Coefficient** | 4 | Genetic Pedigree Coefficient\* | Not Applicable | Ratio | Input |
| **Age** | 5 | Age | Not Applicable | Ratio | Input |
| **BMI** | 6 | BMI | Not Applicable | Ratio | Input |
| **Sex** | 7 | Sex | 0 = Male | Nominal | Input |
| 1 = Female |
| **Pregnancy** | 8 | Pregnancy | 0 = No | Nominal | Input |
| 1 = Yes |
| **Smoking** | 9 | Smoking | 0 = No | Nominal | Input |
| 1 = Yes |
| **Physical\_activity** | 10 | Physical activity  (No. of steps/day) | Not Applicable | Ratio | Input |
| **salt\_content\_in\_the\_diet** | 11 | Salt content in the diet (mg/per day) | Not Applicable | Ratio | Input |
| **alcohol\_consumption\_per\_day** | 12 | Alcohol consumption per day (ml/day) | Not Applicable | Ratio | Input |
| **Level\_of\_Stress** | 13 | Level of Stress  (Cortisol Secretion) | 1 = Less | Ordinal | Input |
| 2 = Normal |
| 3 = High |
| **Chronic\_kidney\_disease** | 14 | Chronic kidney disease | 0 = No | Nominal | Input |
| 1 = Yes |
| **Adrenal\_and\_thyroid\_disorders** | 15 | Adrenal and thyroid disorders | 0 = No | Nominal | Input |
| 1 = Yes |

***\*Genetic Pedigree Coefficient*** *(GPC) of an individual for a particular disease is a continuum between 0 and 1, where*

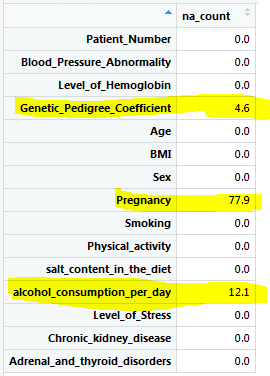
*GPC* ***closer to 0*** *indicates very* ***distant occurrence*** *of that disease in her/his pedigree, and*

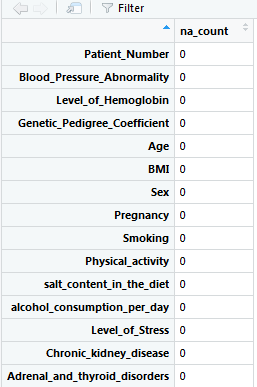
*GPC* ***closer to 1*** *indicates very* ***immediate occurrence*** *of that disease in her/his pedigree]*

**Solution**

1. **Pre-steps:**

The first step is to load the Patient Data into RStudio and do the Exploratory Data Analysis which is the main approach to summarize the characteristics of the data. In this step the most important thing is to clean the data.

1. The first column is Patient Number which looks like a Serial Number. This column will be completely excluded from all the analysis
2. Check all the Data Columns and categorize it into Continuous Variable and Categorical Variables.
3. Continuous Variables:
   1. Level\_of\_Hemoglobin
   2. Genetic\_Pedigree\_Coefficient
   3. Age
   4. BMI
   5. Physical\_activity
   6. salt\_content\_in\_the\_diet
   7. alcohol\_consumption\_per\_day
4. Categorical Variables:
   1. Blood\_Pressure\_Abnormality
   2. Sex
   3. Pregnancy
   4. Smoking
   5. Level\_Of\_Stress
   6. Chronic\_kidney\_disease
   7. Adrenal\_and\_thyroid\_disorders
5. Looking into the data, it has been identified that some of the columns have the NULL values, which needs to be clean as part of Data Cleaning Technique. Below charts says the Percentage of NULL values per 2000 Records.
6. 
7. The chart says,
   1. Genetic\_Pedigree\_Cofficient = 4.6 % NULL Count
   2. Pregnancy = 77.9% Null Count
   3. Alcohol\_Consumption = 12.1 % NULL Count
8. Genetic Pedigree is an important Column for doing the analysis on the Target Variable this can be identified using the Correlation Matrix as per given below.
9. **Treatment for NULL Columns:**
   1. **Genetic Pedigree Coefficient** – It is having 4.6% NULL Count. Either we can exclude all the rows having Null with Genetic Pedigree Coefficient value or we can replace it with some values. In my solution, I have retrieved the Mean Value of the said column and replaced all the Null values with it.
   2. **Pregnancy** – It is having 77.9% NULL values, in my understanding we can replace these values with 0 and can consider there is no Pregnancy in this case.
   3. **Alcohol\_Consumption** – It is having 12.1% NULL Count, in my understanding we can replace these values with 0 and can consider the person is a Teetotaler.
10. After NULL Columns Treatment: below is the updated chart, which shows more clean data.



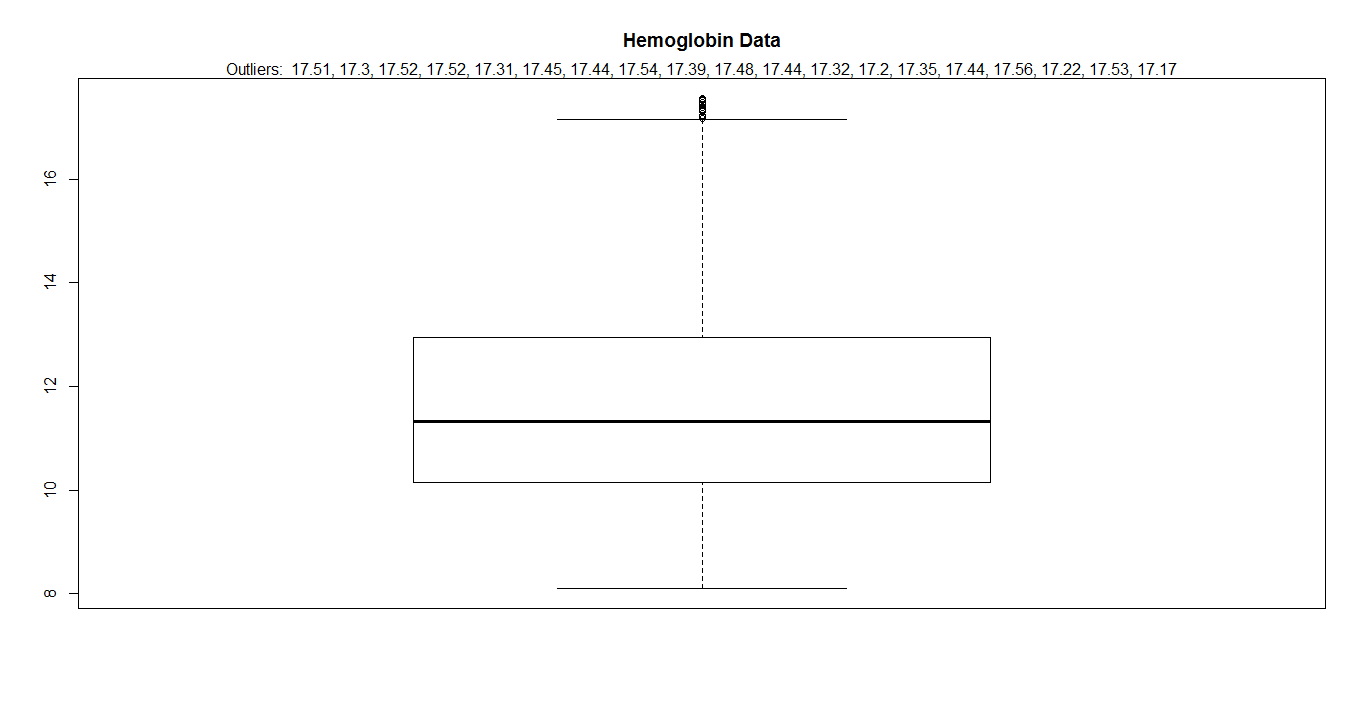
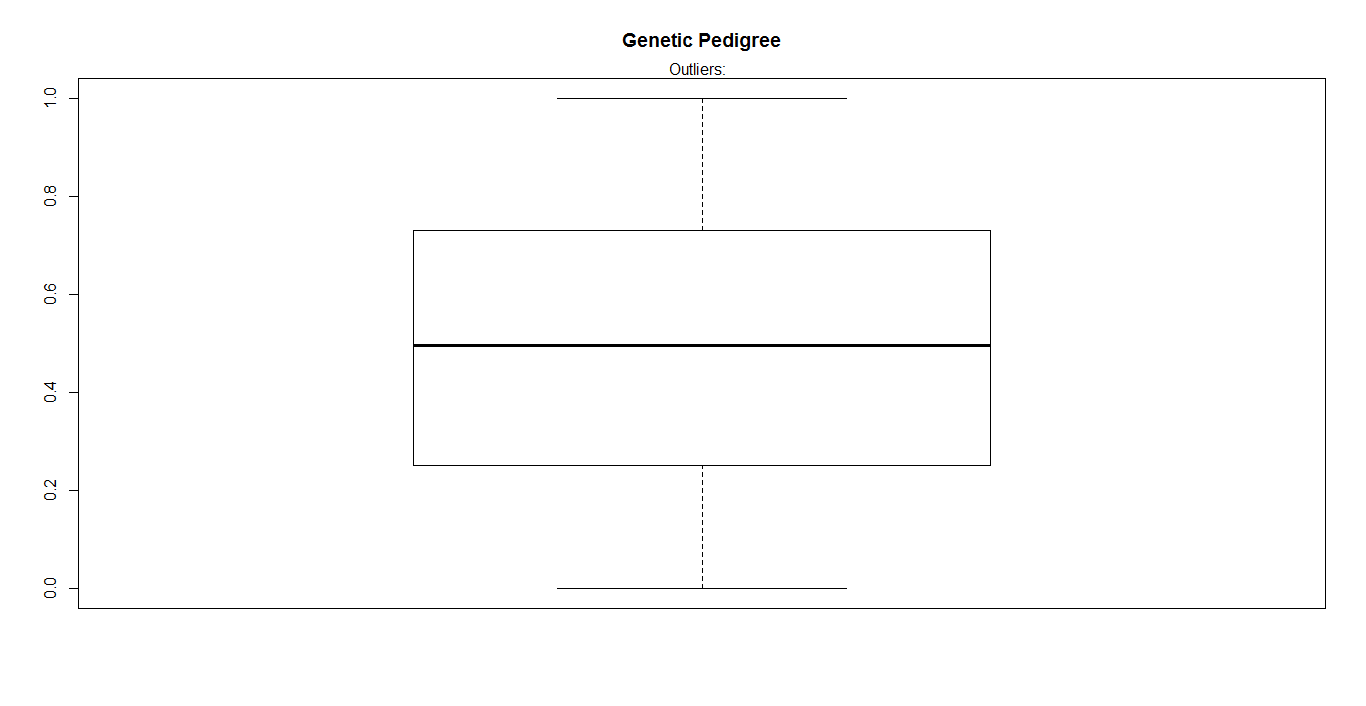
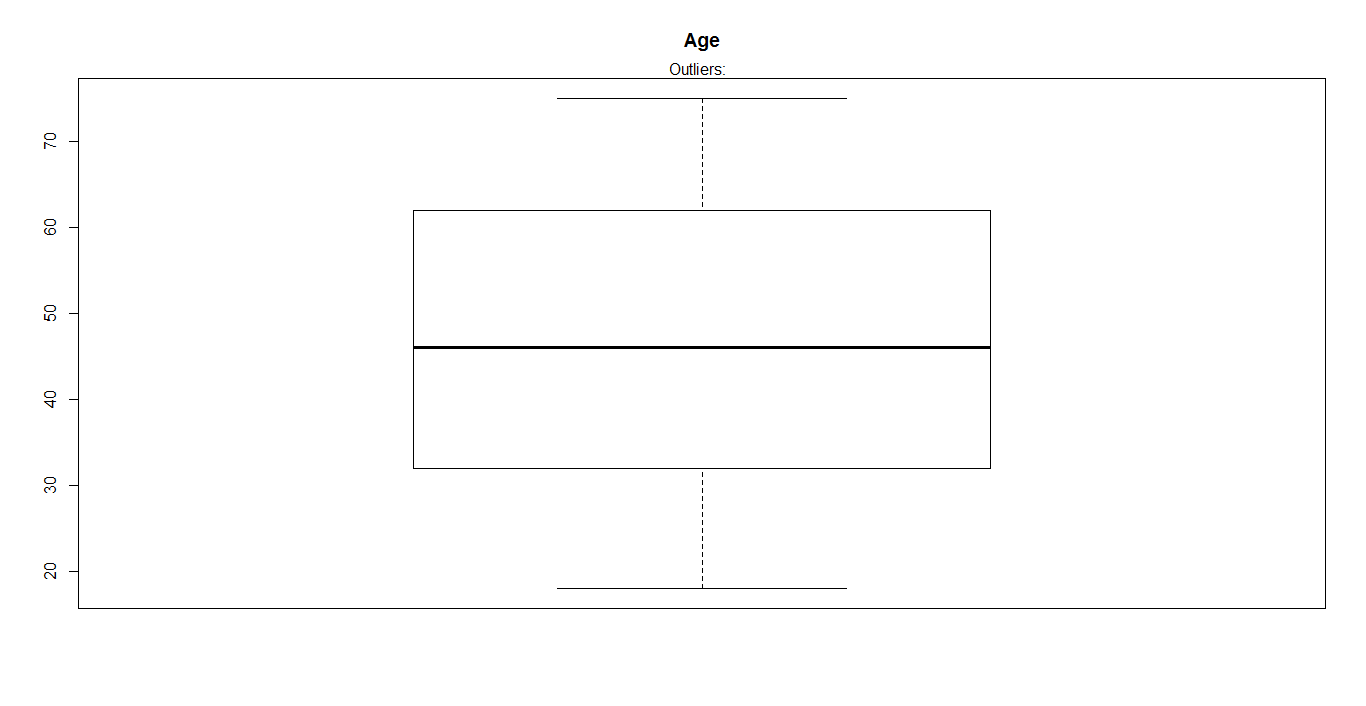
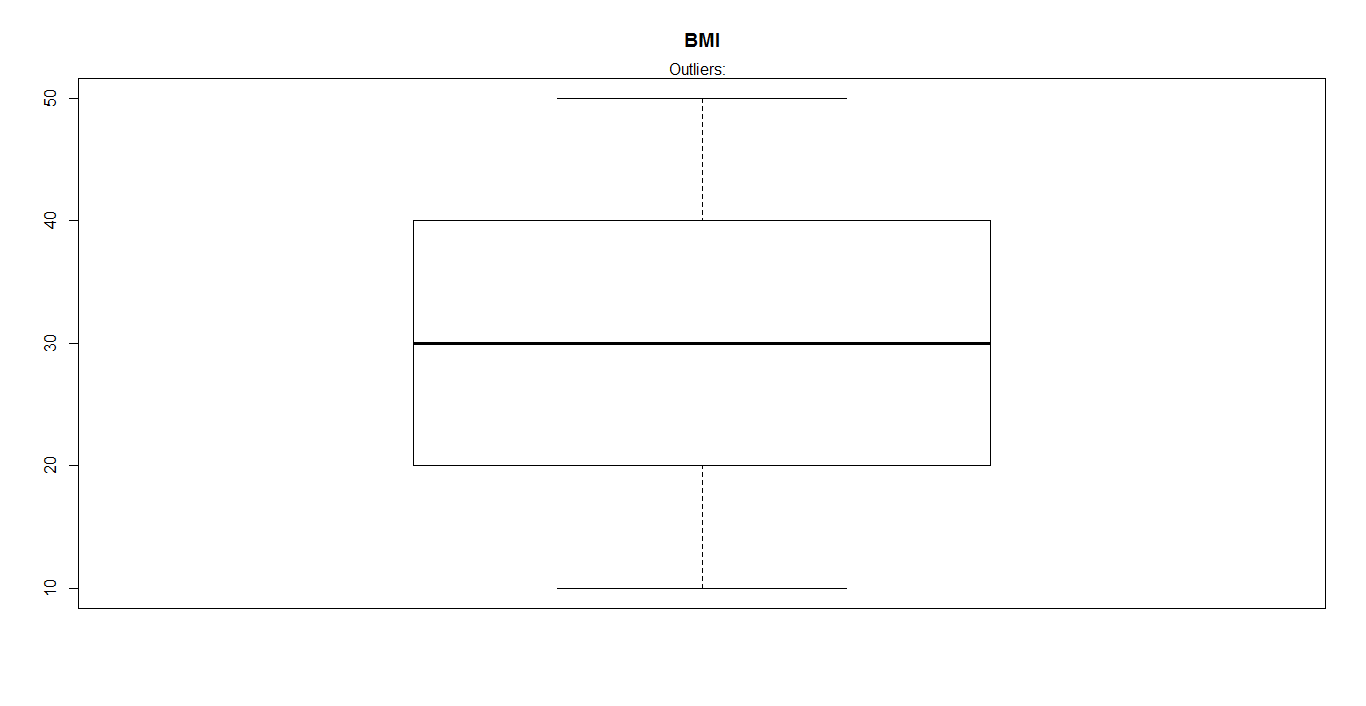
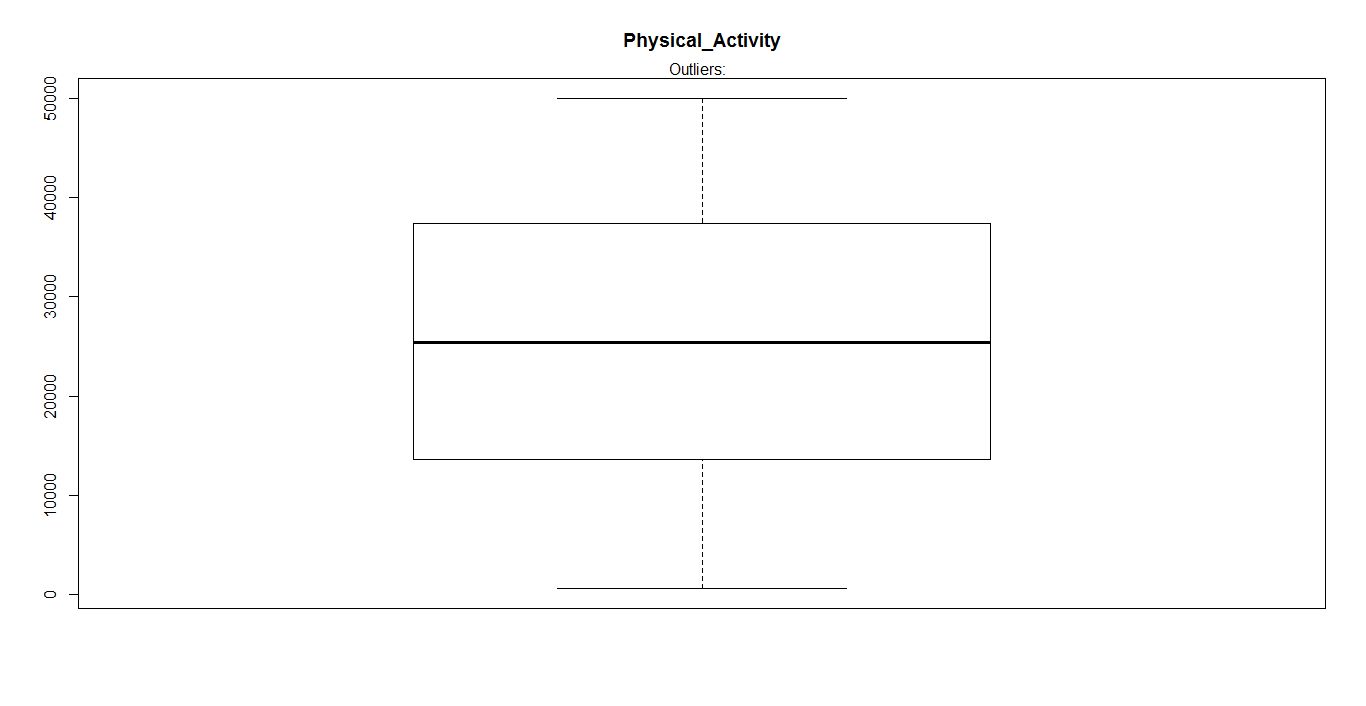
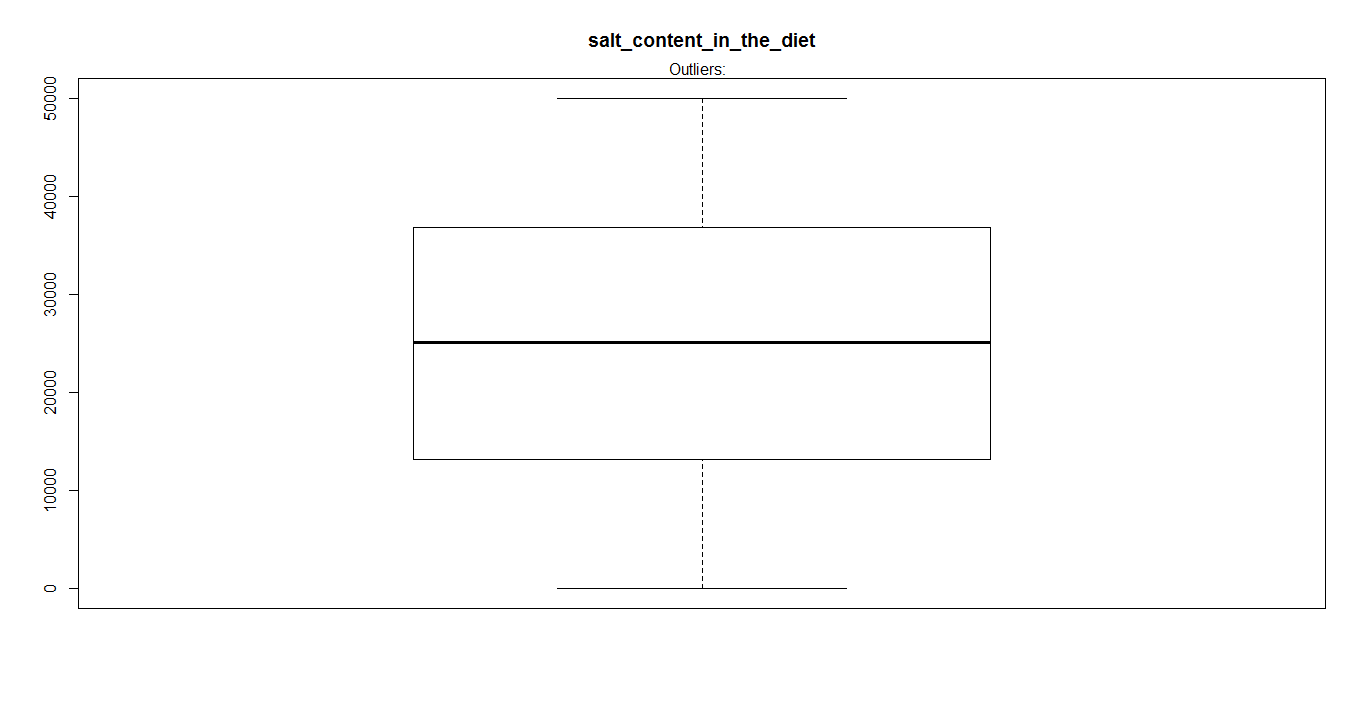
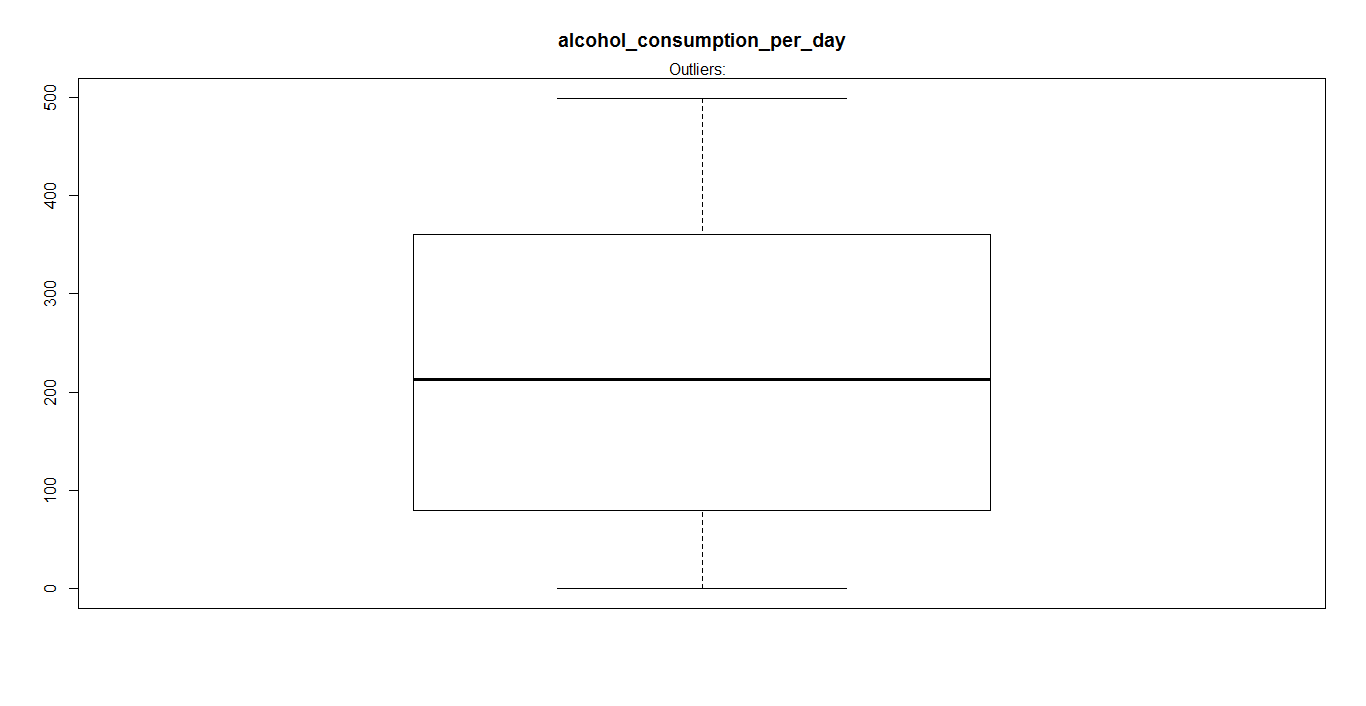
1. **Treatment for Outliers:** The next step is to check if there are any outliers in the data. Outliers treatment are really an important step, otherwise it can distort the data and affect the Prediction.
   1. **Hemoglobin Outliers Check:**

The normal range for hemoglobin is:

For men, 13.5 to 17.5 grams per deciliter

For women, 12.0 to 15.5 grams per deciliter

In our case, it is showing some outliers for Hemoglobin – but those outliers are still closely related to the Acceptable Limits, hence, we are not going to exclude any values from our analysis.

* 1. 
  2. **Genetic Pedigree Outliers**: None Identified 
  3. **Age Outlier:** None Identified 
  4. **BMI Outlier**: None Identified 
  5. **Physical Activity Outlier:** None Identified 
  6. **Salt Content Outlier**: None Identified 
  7. Alcohol Consumption Outlier: 

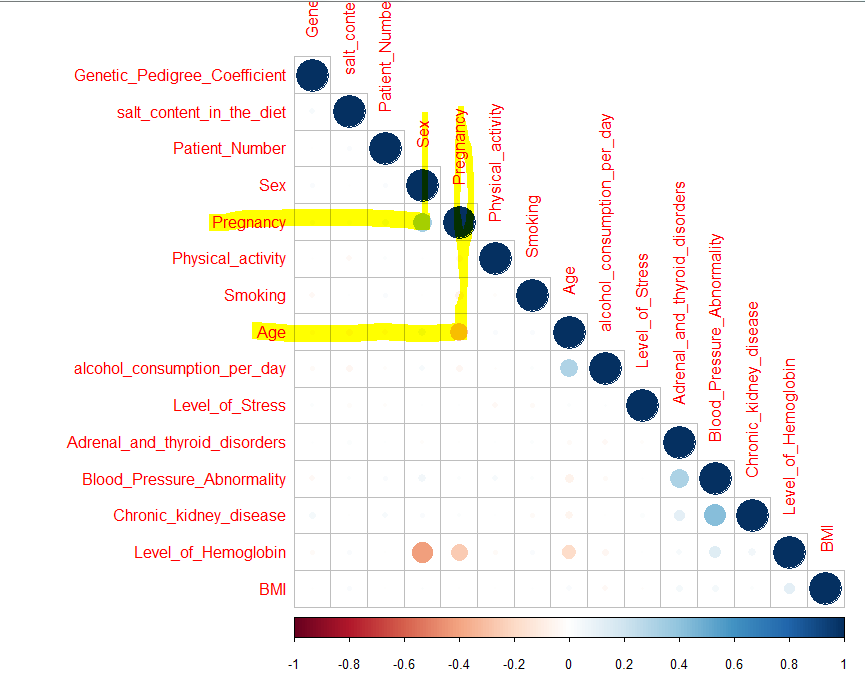
1. **Check for the Target Variable Response**: In the current Data, we have count of:

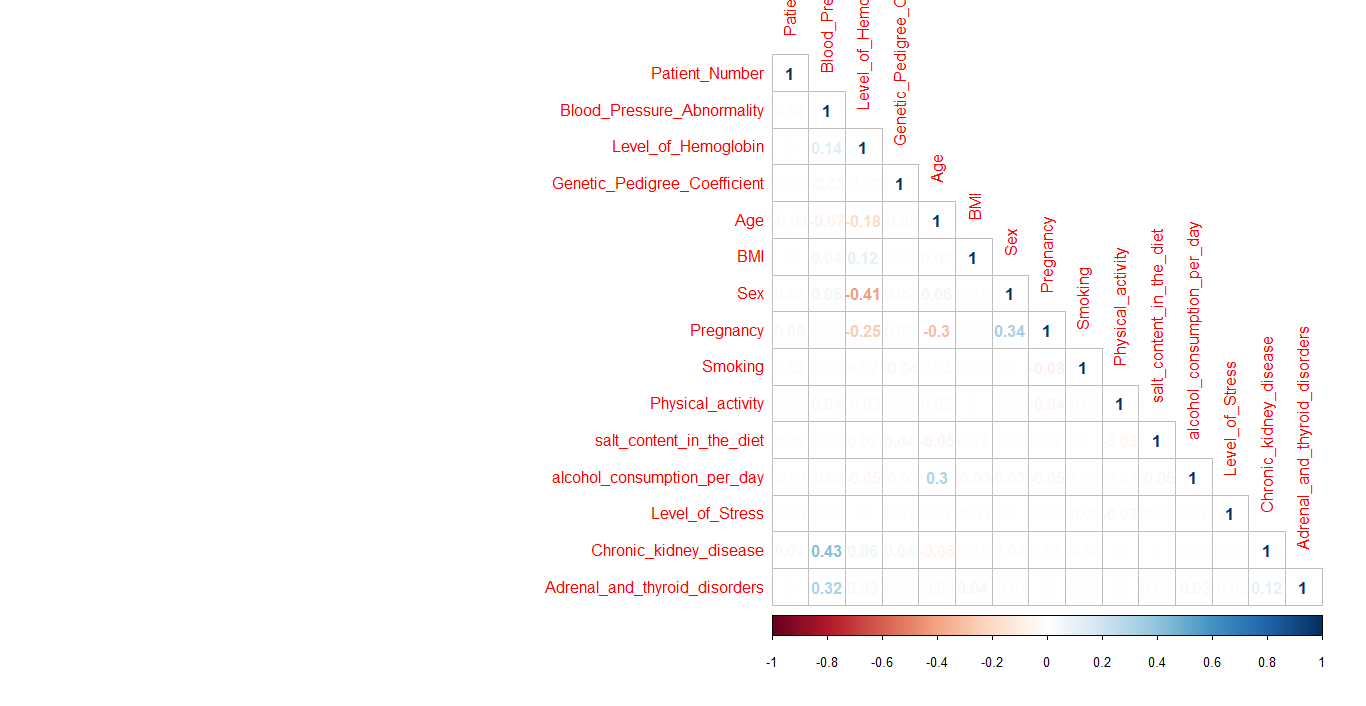
Normal B.P = 50.65% and Abnormal B.P = 49.35%.

1. **Correlation Matrix of the Variables/Columns:** The data says that there is very less co-relation of the variables with each other. But this step is really important to understand what all variables can make an impact on others. Few examples of Highly co-related variables below.

**e.g. 🡪 Pregnancy is highly related with Gender.** This has been highlighted inside the figure with yellow mark.

🡪 **Age with Pregnancy** This has been highlighted inside the figure with yellow mark.

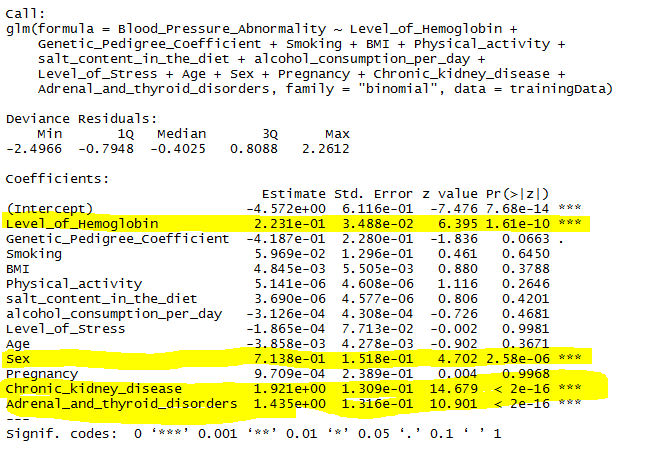


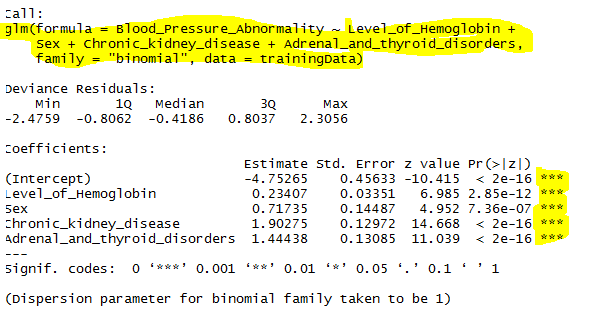


1. **Building ML Models – Using All Columns.**

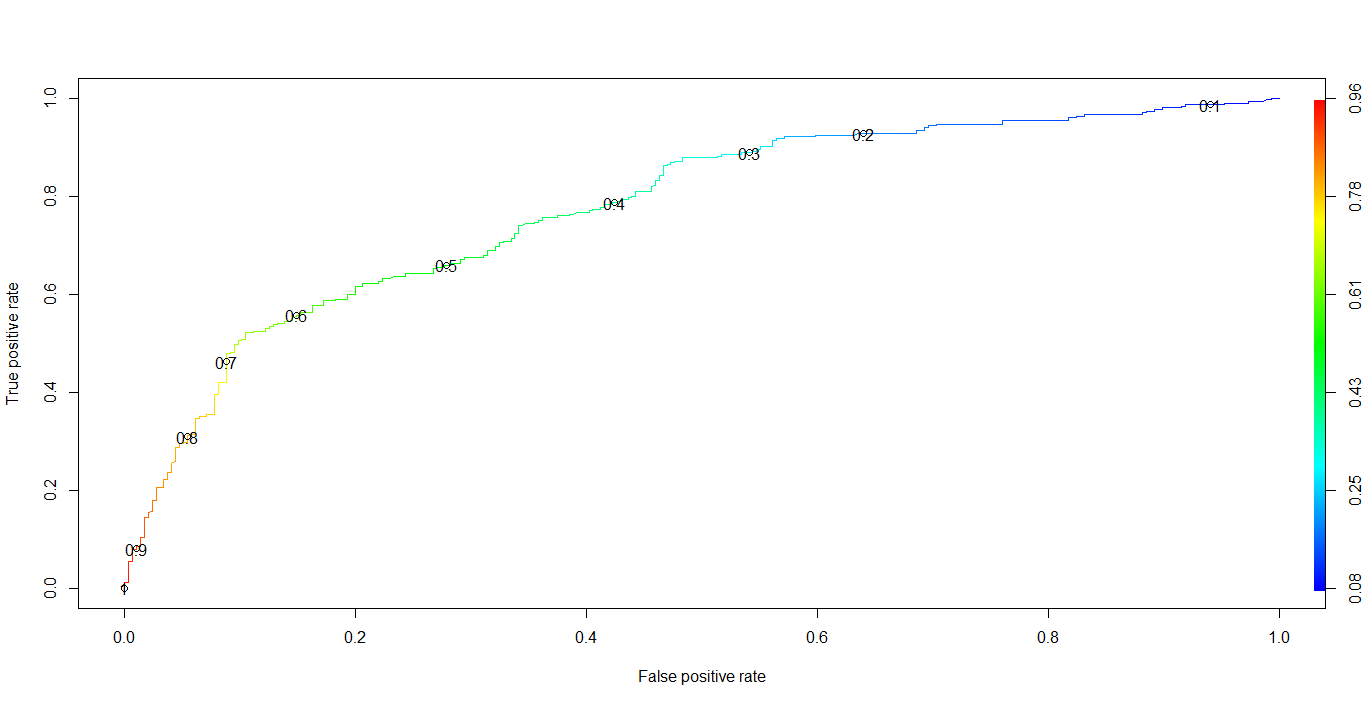
**Pre-Steps:**

As per the requirement, we need to implement the Models considering Hemoglobin Level and Genetic Pedigree Coefficient. On the very start of developing below models, I have divided my data set into 2 parts:

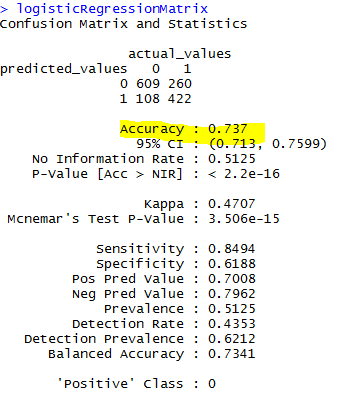
1. Training DataSet – 70% of the Records given will be used for training the model.
2. Testing DataSet – Remaining 30% of the Records given will be used for testing the model.
3. The above dataset will be ingested to all the models in our analysis.
4. ***Building Logistic Regression Model and Check Accuracy:***
5. Build a Logistic Regression model using only 2 Variables as defined above.
6. Library used: ROCR
7. 
8. If we look at the summary of the model, it has been identified that the Hemoglobin Level is the most significant input for the model as its p-value was less that .05.
9. Other Variables which are important like: Sex, Chronical Kidney Disease, Thyroid Disorders.
10. Rest of the variables, can be disregarded as there is no importance to it.



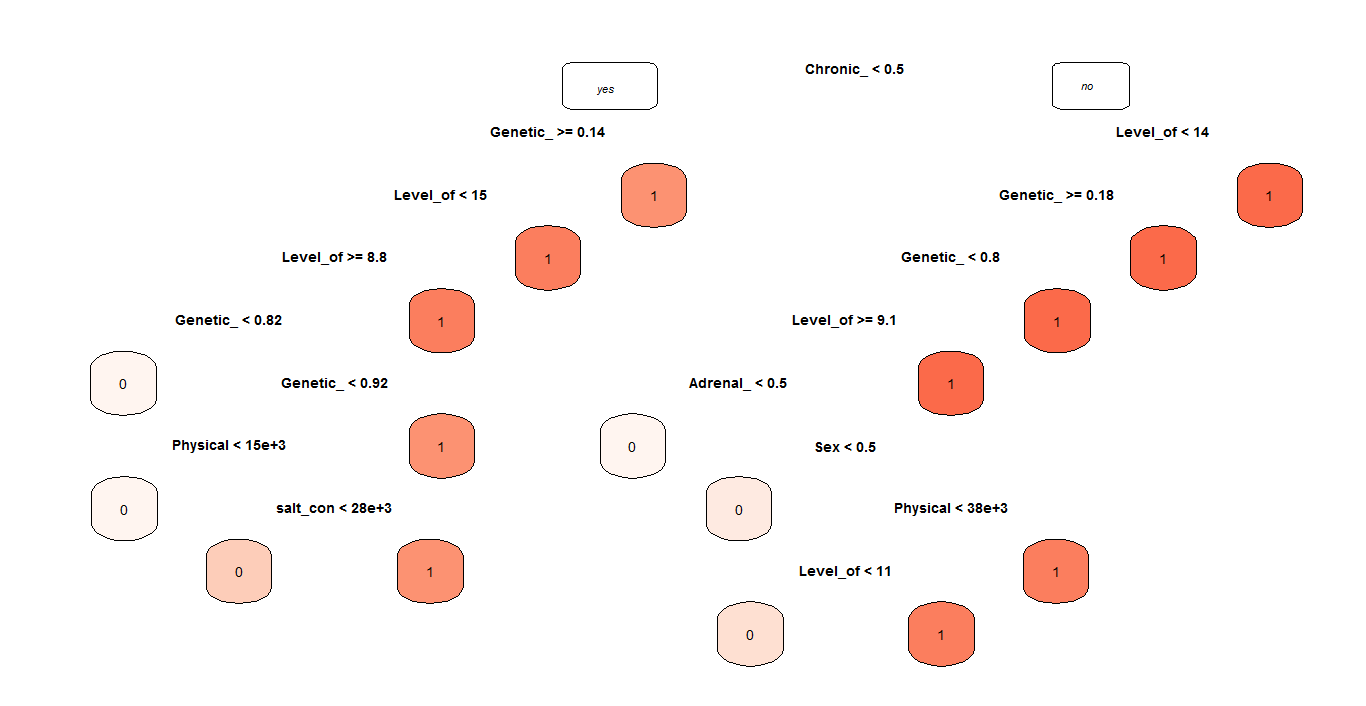
1. Next, step is to predict the data with the Testing Data Set.
2. Created the ROC Curve to see the predicted results which should be in between 0.1 to 1.0. On Y Axis – True Positive Rate, on X-Axis False Positive Rates.



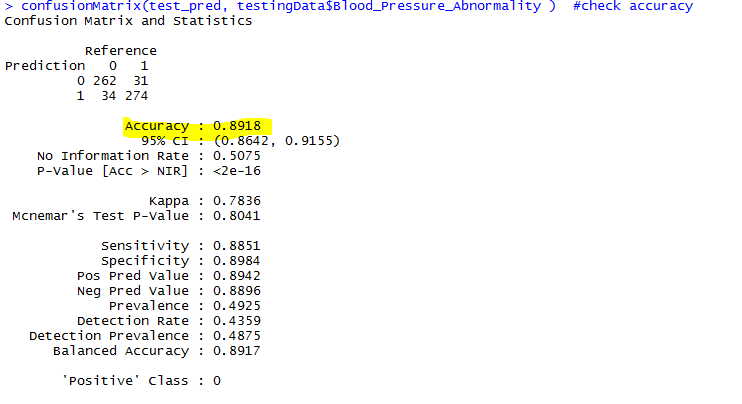
1. Clearly, my ROC Curve is getting bended from 0.5, so I took the threshold of 0.5 and put it into the Prediction results.
2. Generated the Confusion Matrix.



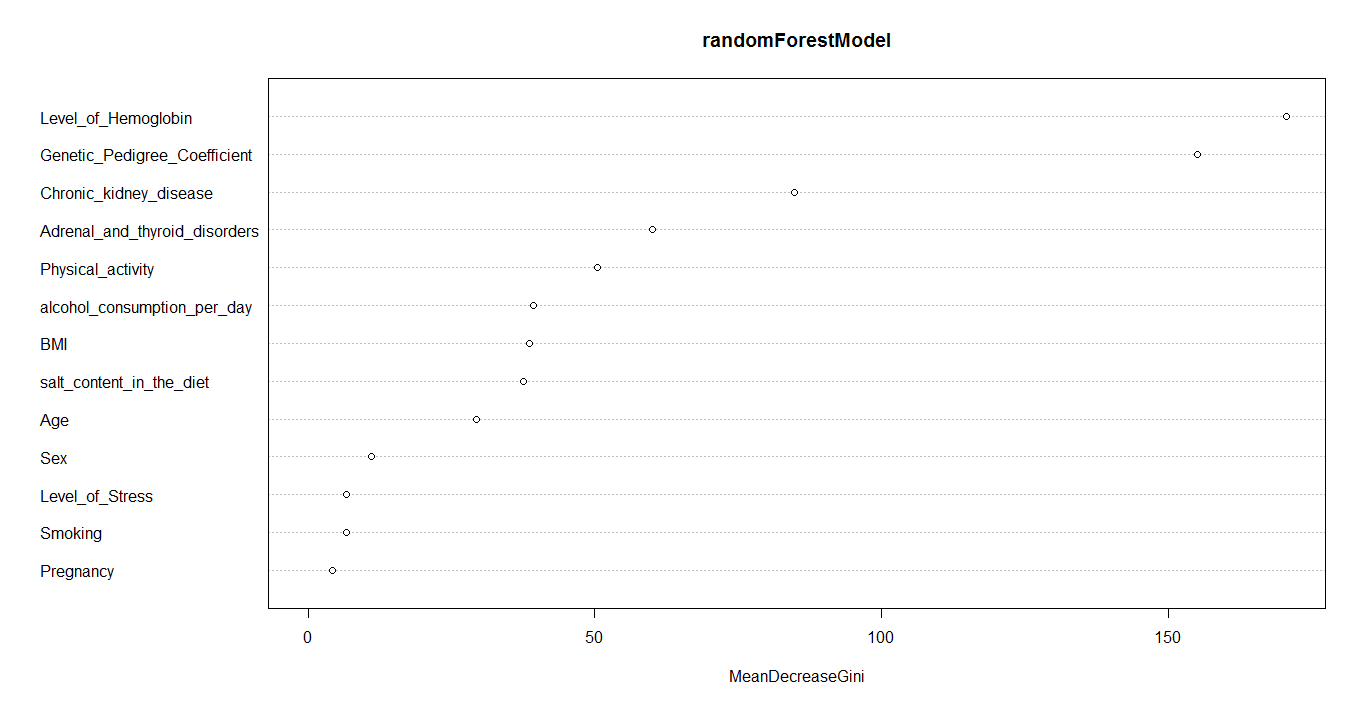
1. Output: 73.7% Accuracy
2. ***Building Decision Tree Model and Check Accuracy:***
3. Build the Decision Tree model and generated the Tree using **prp** method.
4. Library used:
   1. rpart
   2. rpart.plot
   3. caret

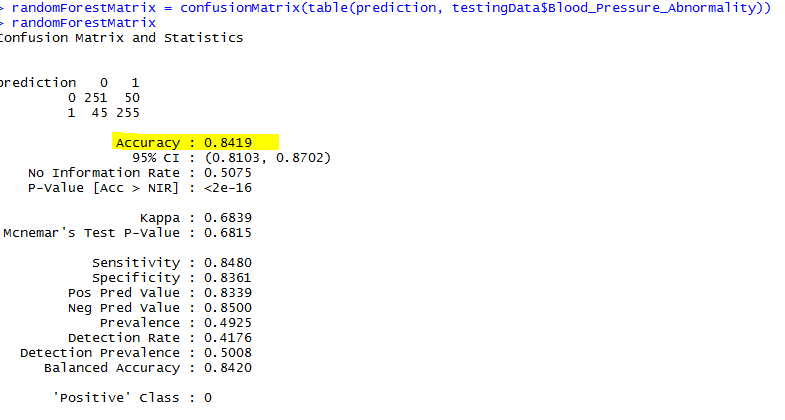


1. Checking the Accuracy of the Model with the Testing Data:

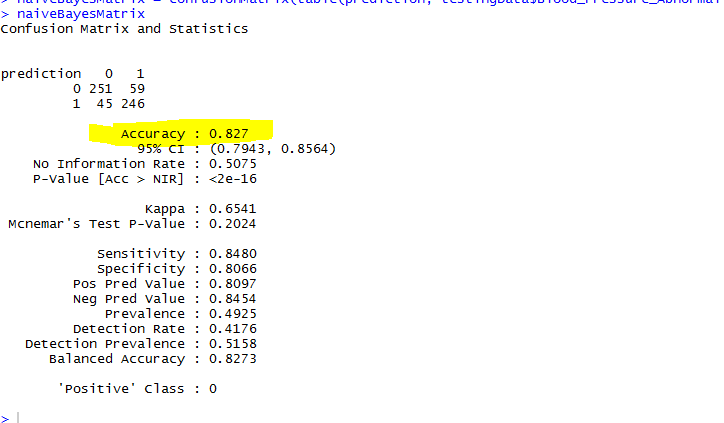


1. Output: 89.18 % Accuracy
2. ***Building Random Forest Model and Check Accuracy:***
3. Build the Random Forest Model.
4. Library Used: **RandomForest**
5. In this model, we check for the Important Variables and in my study the important variables are Hemoglobin Level as per below figure:



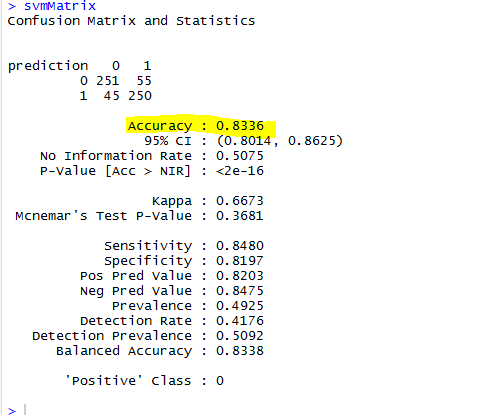


1. Output: 84.19 % Accuracy
2. ***Building Naïve Baye’s Model and Check Accuracy:***
   1. Libraries Used in building this model: E1071



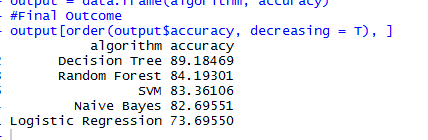
* 1. Output: 82.7 % Accuracy

1. ***Building Support Vector Matrix Model and Check Accuracy:***
   1. Libraries Used in building this model:
      1. E1071



* 1. Output: 83.3 % Accuracy

1. **Model Summary and Final Output with Multiple Variables in Decreasing Order.**



The output says, that Decision Tree Model is giving the high Accuracy with the current set of data and Input Variables.