## **Healthcare Management Information and Analytics**

## **Assignment 3: Screening for Chronic Kidney Disease**

By, Naveen L Deevi, 675930999

# **Introduction:**

This case study is about finding out people who has high probability of contracting to Chronic Kidney Diseases. So, to identify likely patients, I was provided with demographic and health related information of the people to be examined. The dataset that I was provided with, has 6000 data points(people) with 33 variables(observations on different fronts i.e., demographic, health and lifestyle related information).

Task ahead of us is to do a detailed data analysis and produce an algorithm that can predict the chances of a person to be infected by chronic kidney disease.

**Methodology:**

We divided the data analysis part into 5 parts as following:

1. Descriptive Statistics
2. Handling sparsity of the data
3. Variables Selection
4. Model Selection
5. Model application on Validation data set.

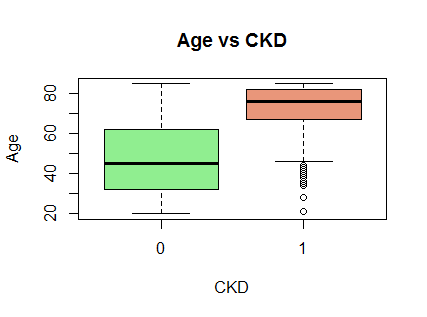
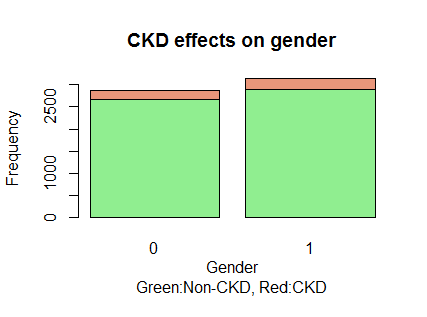
# **Descriptive Statistics:**

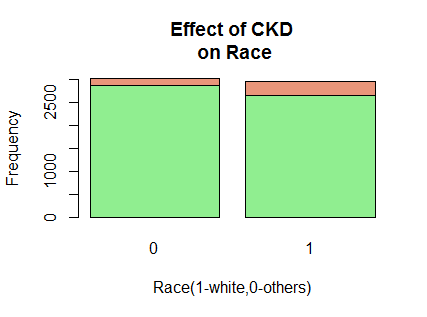
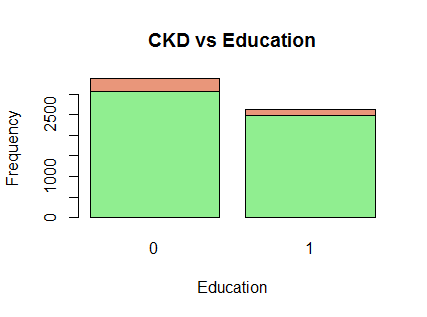
The data contains following variables:

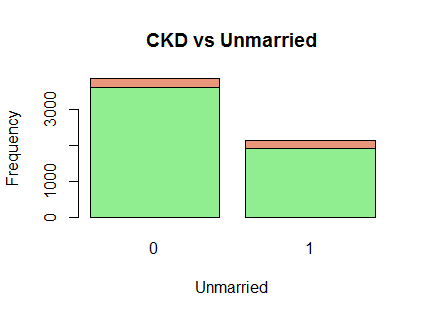
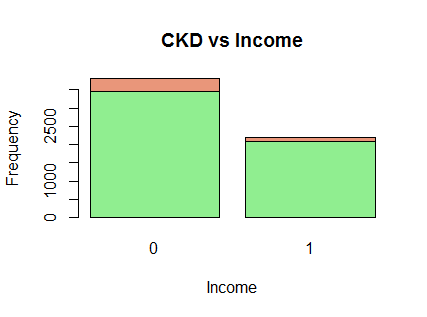
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Type** |  | **Variable Name** | **Type** |
| ID | Numerical |  | LDL | Numerical |
| Age | Numerical |  | Total Chol | Numerical |
| Female | Binomial |  | Dyslipidemia | Binomial |
| Racegrp | Categorical |  | PVD | Binomial |
| Educ | Binomial |  | Activity | Categorical |
| Unmarried | Binomial |  | Poor Vision | Binomial |
| Income | Binomial |  | Smoker | Binomial |
| CareSource | Categorical |  | Hypertension | Binomial |
| Insured | Binomial |  | Fam Hypertension | Binomial |
| Weight | Numerical |  | Diabetes | Binomial |
| Height | Numerical |  | Fam Diabetes | Binomial |
| BMI | Numerical |  | Stroke | Binomial |
| Obese | Binomial |  | CVD | Binomial |
| Waist | Numerical |  | Fam CVD | Binomial |
| SBP | Numerical |  | CHF | Binomial |
| DBP | Numerical |  | Anemia | Binomial |
| HDL | Numerical |  | CKD(Dependent) | Binomial |

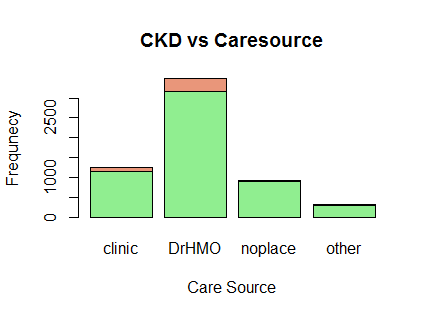
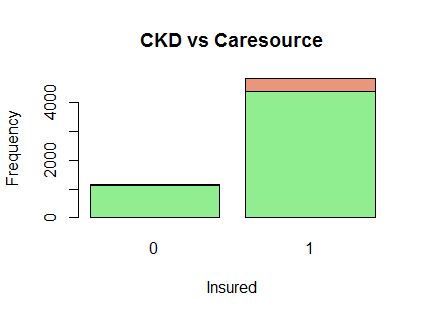
To understand the distribution and the effects of Independent variables on dependent variables I did some descriptive statistical operations on the data. And found out some are impacting the dependent variables positively or negatively and some are not having any effect on dependent variables. Though we cannot confirm the correlation between variables without statistical substantiation, I believe these statistics would give a comprehensive understanding of data.

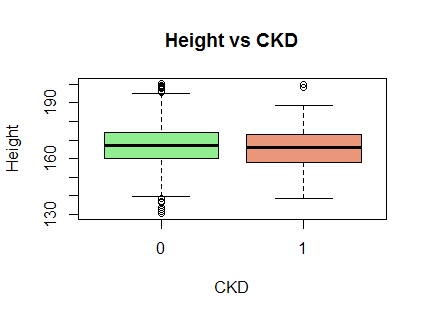
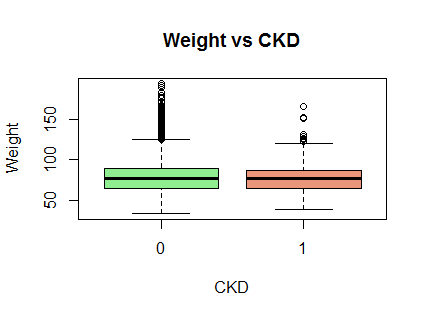
Descriptive Statistics of Variables and CKD: (Green: Non CKD, Red: CKD)

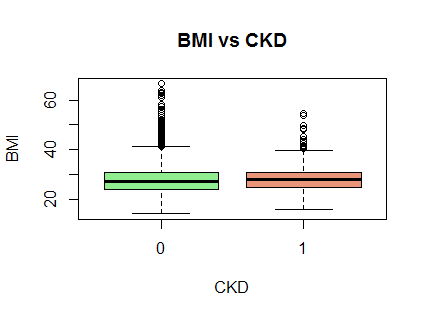
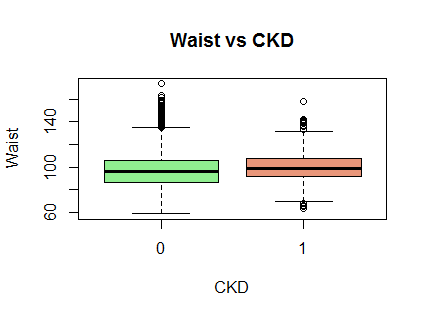
 

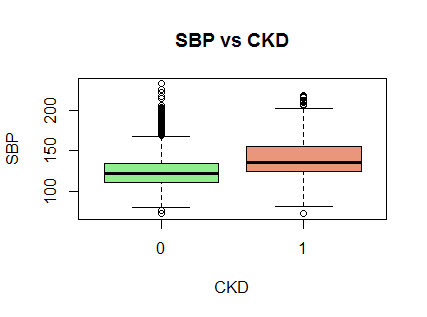
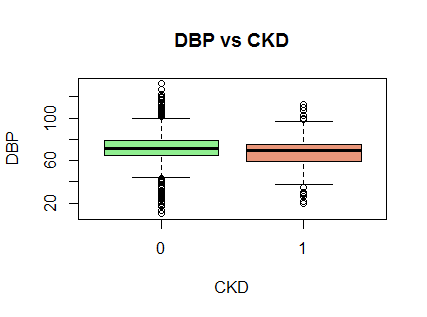
 

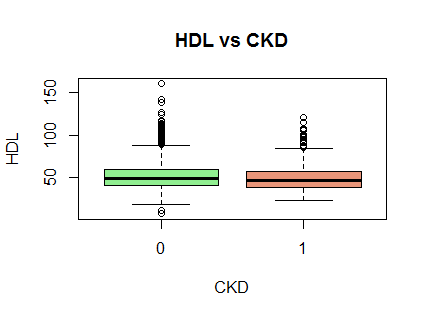
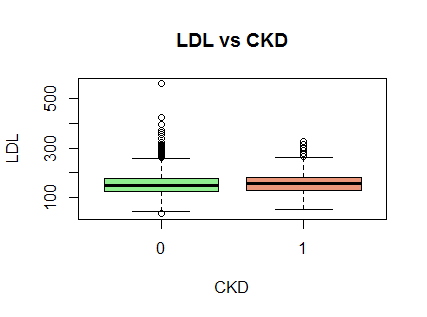
 

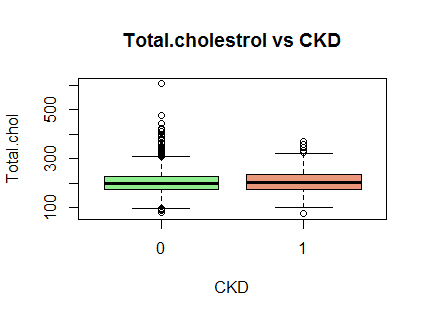
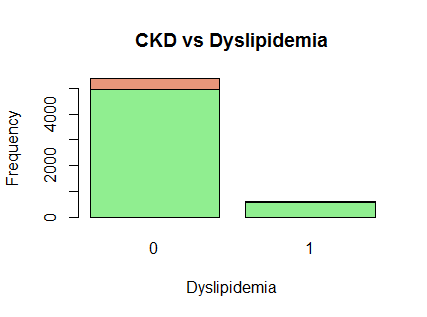
 

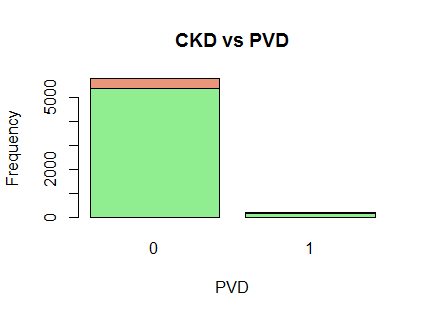
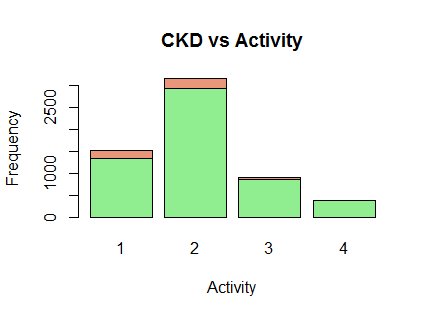
 

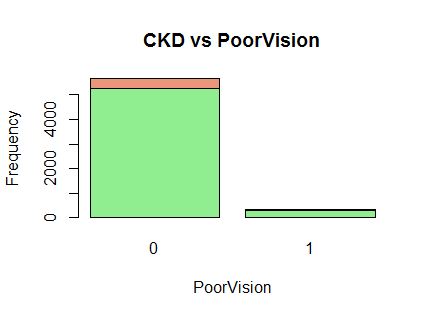
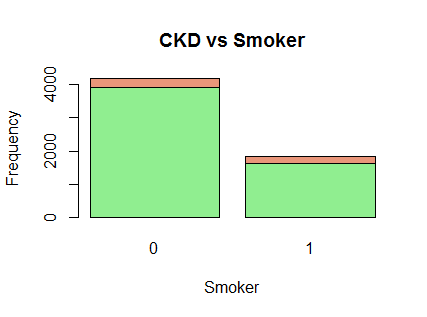
 

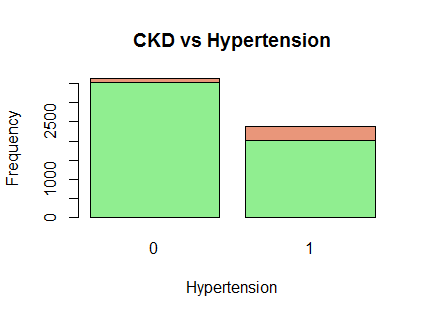
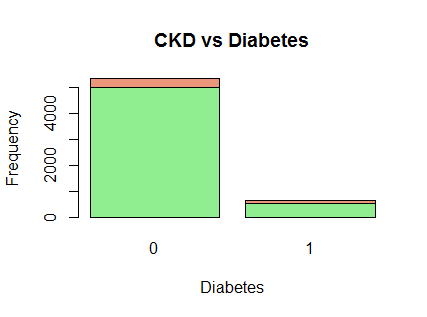
 

From the above graphs I was able to understand the distribution of the variables and how they effect the target variable.

# **Handling sparsity of Data & Variable Transformation:**

Handling missing data is a major challenge in any data analysis. Missing values shoukd be handled very carefully, if they are not handled carefully they might change the distribution of the variables and that change the course of the data analysis. Using Domain knowledge is the best way to handle missing data or advanced techniques as MLE or Multiple imputation can be used to handle missing data. In Healthcare industry any datapoint which has many missing values it is advisible to remove that data point. And care hasa to be taken that the distribution of the variable is not effected much.

Categorical Variables: In this type of variables, a dominant type is placed in as missing value. Education, Unmarried, Income, Insured, Activity, Poor Vision, Hypertension, Stroke, CVD, Fam.CVD, CHF, Diabetes and Anemia are the categorical variables that are handled as above mentioned way.

Numerical Variables: The mean of the variable is placed in as missing value in these type of variables. Weight, Height, BMI, Waist, SBP, DBP, HDL, LDL, Total.Cholestrol are handled with mean replacement.

**Variable Transformation:**

As in the case study it is mentioned that people of White race are more susceptible to CKD, the variable Race is transformed to binary variable of Whites and others. White is taken as base line, that is white would be 1 and others would be 0.

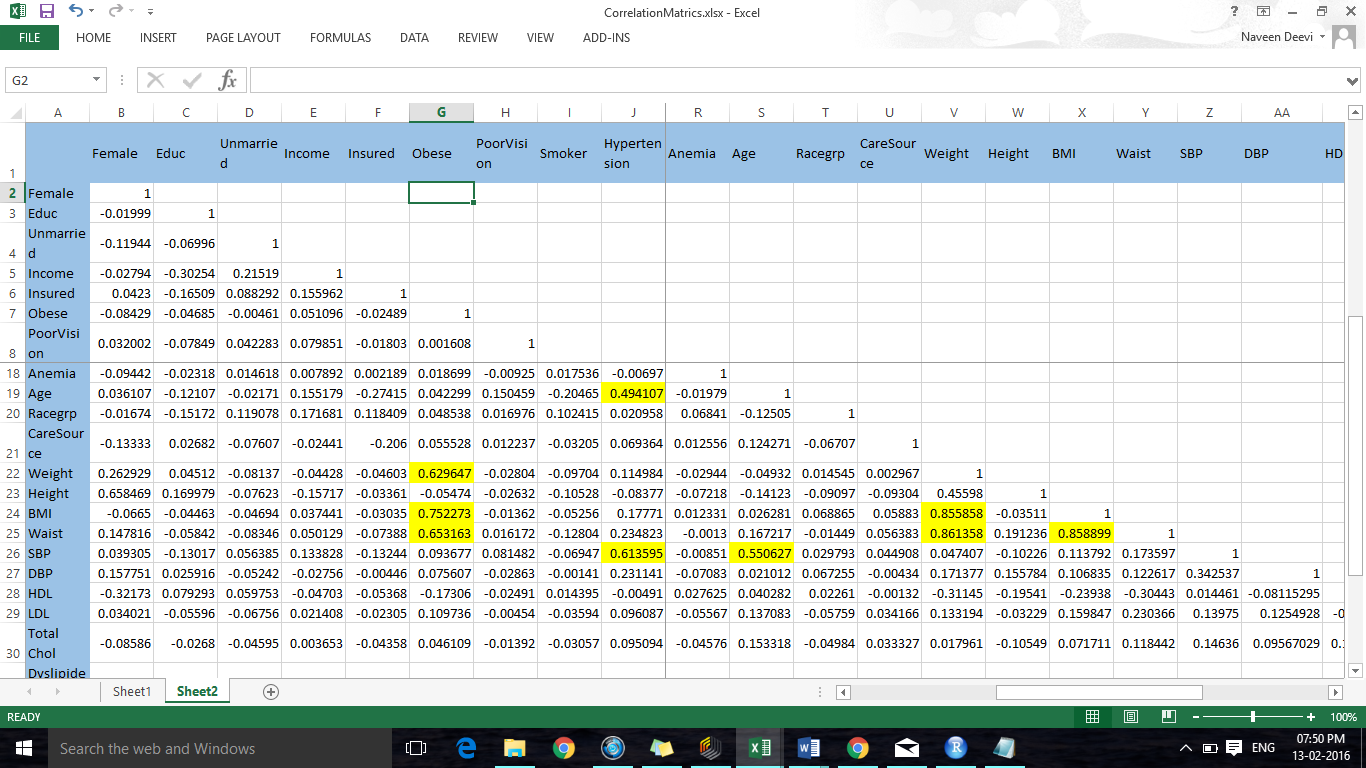
Variable Obese is calculated from BMI hence, as stated in case study any data point with BMI more than 30 is considered obese and others are not.

**Data Split:** In 6000 data points I used 4000 to train the model and 2000 for testing the algorithm.

# **Variable Selection:**

In any data analysis there will be numerous number of variables, but all variables are not important for predicting outcome variable. Variables which effect the prediction should only be considered for data analysis, and other variables should be discarded, otherwise they will affect the prediction in negative way. So task of variable selection is to select variables that can explain the dependent variable.

Multicollinearity: Multicollinearity is undesirable in any data analysis. Multicollinearity is nothing but one predictor variable is dependent on another predictor variable. In other words the deviation of one independent variable is explained by another independent variable which makes them both correlated and not independent. So, first a correlation matrix should be produced and based on that correlated variables should be eliminated for further analysis.



In the above matrix yellow marked are the variables that are correlated with each other. The threshold I selected for this is +0.5 and -0.5. After correlated variables are eliminated we can move for further variables selection processing.

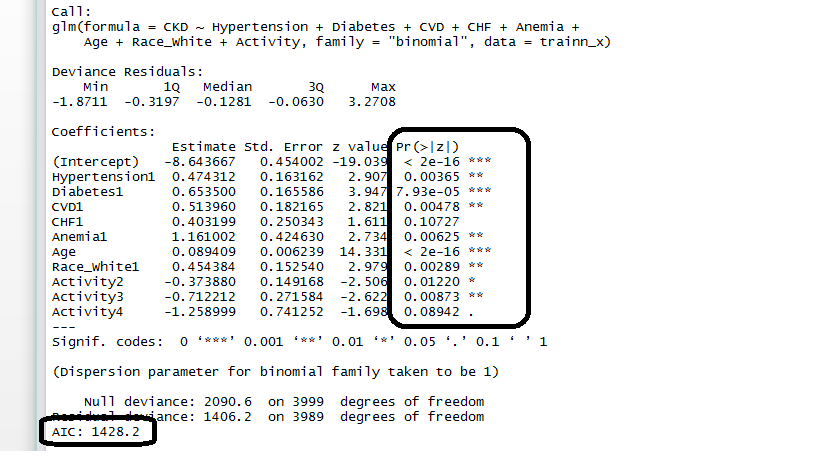
**Forward Selection Process:**

In this process I selected one variable at a time and found out the deviation that a variable explains in target variable. If a variable explains substantial deviation in target variable I will select that variable for further processing otherwise I will drop it. I used P value and AIC to do this filtering. If **P value** is less than 0.05 then it is most important variable. **AIC- Akaike information criteria** is a metric used to compare the models, low value of AIC is desirable.

From this selection process I filtered out the most important variables, those are:

Hypertension, Diabetes, CVD, CHF, Anemia, Age, Race\_White, Activity.

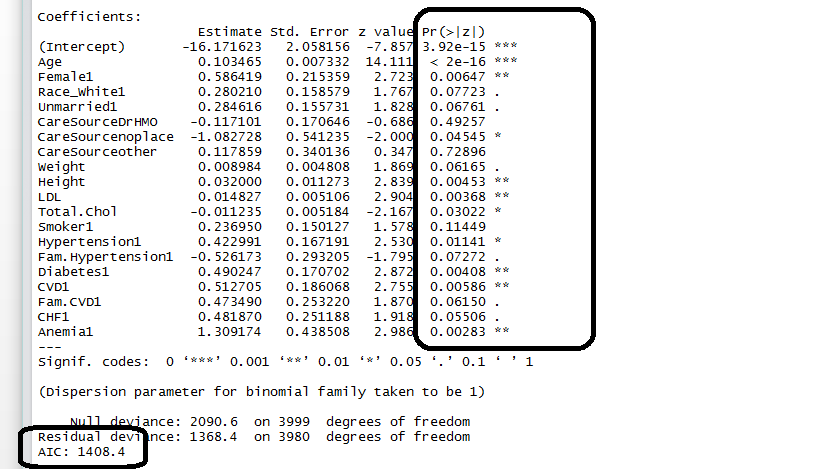
R Output of importance of these variables is:



**Backward Elimination process:**

Some variables will perform well when they are combined in the prediction. They may not perform as expected if they are applied individually, so this kind of variables will not be captured in forward selection process. So I have applied backward selection process where all the variables are used together to predict target variable, then I kept on dropping one variable at a time which are not valuable and not useful in analysis. Here as well I used P value and AIC to select the variables.

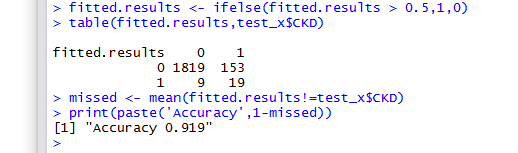
Following are the outputs of my backward elimination process.



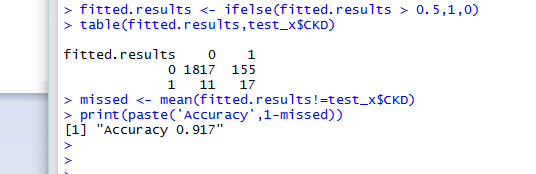
# **Model Selection:**

From the above two processes I have selected 2 models. Forward selection process with AIC 1428 and Backward elimination process with AIC 1408. Now on applying both on the testing data I would select the model which will suit my data most.

Froward Selection Process Output on testing data:



Backward Elimination process output on testing data:



From the above output I am selecting the variables that are out of Forward selection process.

# **Model application on Validation data set:**

I have applied the above selected model (forward selection variables) to the validation data set, which is for the data points with ID’s 6001 to 8819. Out of 2819 validation data set, from my model I identified 44 as 1 and 2775 as 0’s. Following is the csv file having predictions in CKD column for ID 6001 to 8819. Linear Discriminant Analysis, Support Vector Machine and KNN models are applied on the data and prediction is made with an accuracy 90%. Attaching R file as well for reference if required.



**By,**  
 **Naveen L Deevi,**

**675930999**