

**CSC 240 – DATA SCIENCE WITH R**

# PROJECT REPORT

|  |  |
| --- | --- |
| **NAME**    **UNIQUE ID**    **YEAR**    **QUARTER**    **DEPARTMENT**    **FACULTY NAME**    **ACADEMIC YEAR** | SK.THANUSHRAAM      E0120040      II      Q1      B.Tech CSE (AI/ML)      Prof. Ramya M      2021 - 2022 |

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SNO** | **TITLE** | **PAGE NUMBER** |
| 1. | **Problem Statement** | **03** |
| 2. | **Objective** | **03** |
| 3. | **Loading the data** | **04** |
| 4. | **Data Exploration** | **05** |
| **5.** | **Data Cleaning** | **08** |
| **6.** | **Data Visualisation** | **11** |
| **7.** | **ML Algorithm** | **16** |
| **8.** | **Result** | **23** |

**CHRONIC KIDNEY DISEASE DATASET ANALYSIS**

## PROBLEM STATEMENT

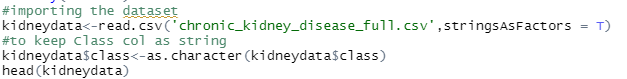
This dataset contains different person’s health report with person affected with chronic kidney disease and those who are not along with their age. The key problem here is to clean the data and use required model to predict whether the person would have chronic kidney disease or not using the given dataset

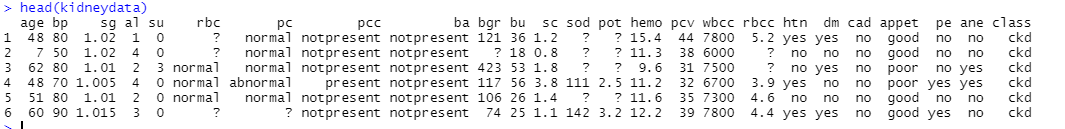
## OBJECTIVE

* To explore and understand each attribute of the dataset.
* To clean the data and dropping unwanted columns and replacing mistyped values
* To select appropriate prediction model algorithm that combines several features to predict whether the person would have chronic kidney disease or not
* To visualise and understand the relationship between each attribute using various graphs and also how they would affect the target variable

**DATA LOADING**

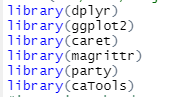
We have kept target column(Class) as character type in order to use them in model





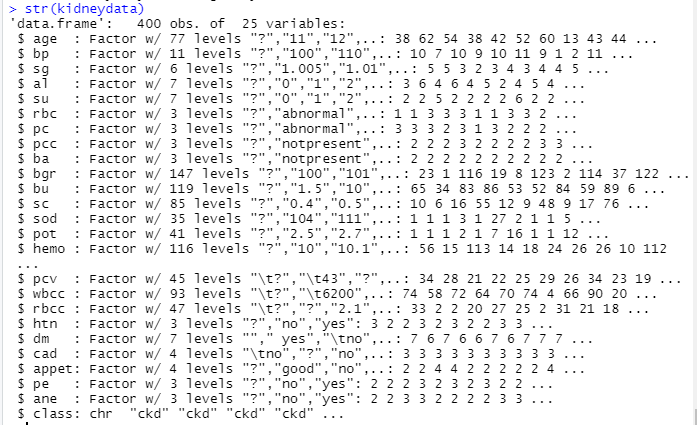
This dataset is taken from UCI Machine Learning Repository.

**IMPORTING THE LIBRARIES**

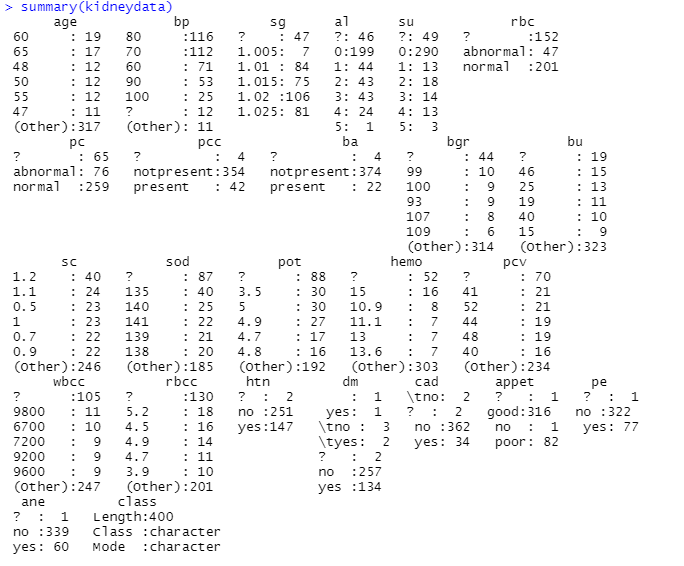
****

**DATA EXPLORATION**

Structure of the given dataset.

****

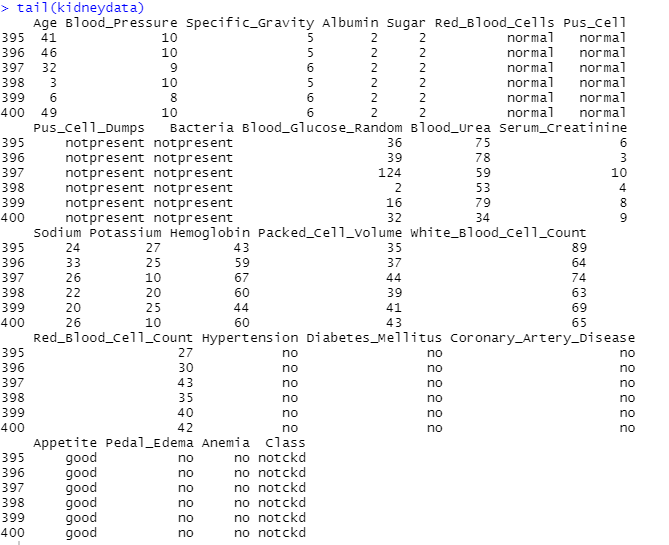
Summary of the given dataset.



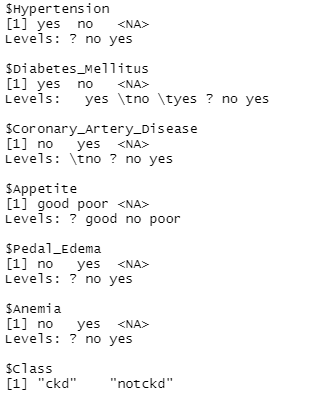
Names of all the column present in the data

****

Displaying last 5 columns of data using TAIL function



Checking unique values of the column



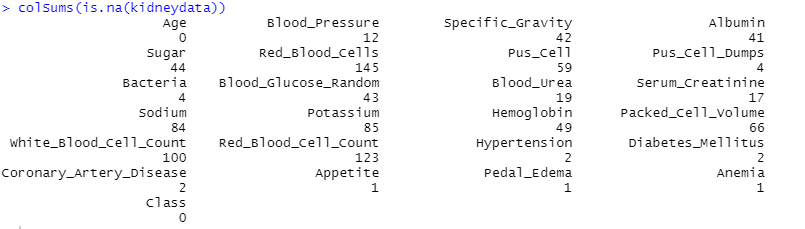
**DATA CLEANING**

As we can see from the dataset there are many wrongly entered data and missing values. We will mutate each and every row so that our data is clean.

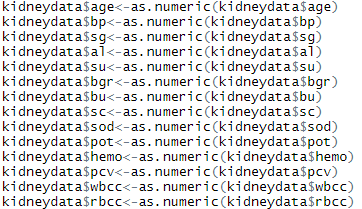
Replacing “?” with NA so that to compute total NA values in the dataset.



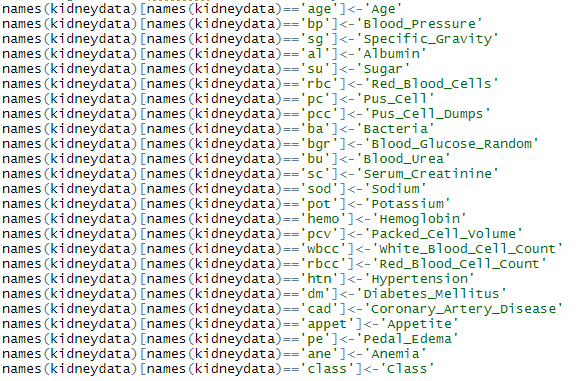
You can see total NA values in each column



Converting suitable character column to numeric column to perform mathematical operations on them



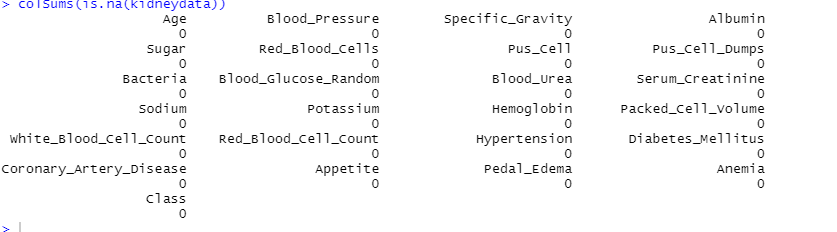
In order to make the columns to be read easier, we have renamed them



Replacing the NA values with mean or median depending on the column. Categorical columns are renamed accordingly.

****

After cleaning the data and replaced all the missing values

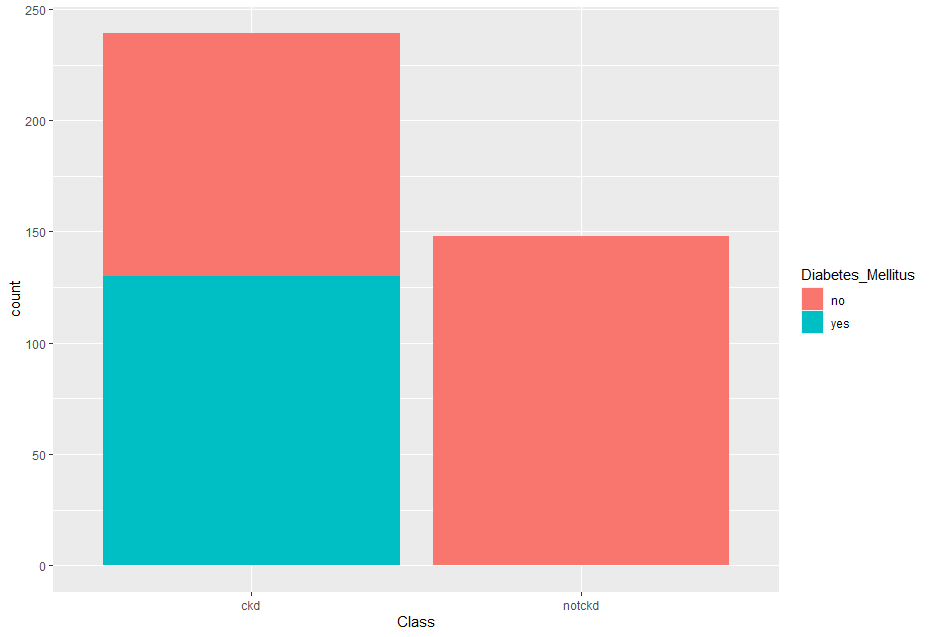
****

**DATA VISUALISATION**

**Bar Graph**

Comparison of people with Diabetes Mellitus and CKD and without CKD



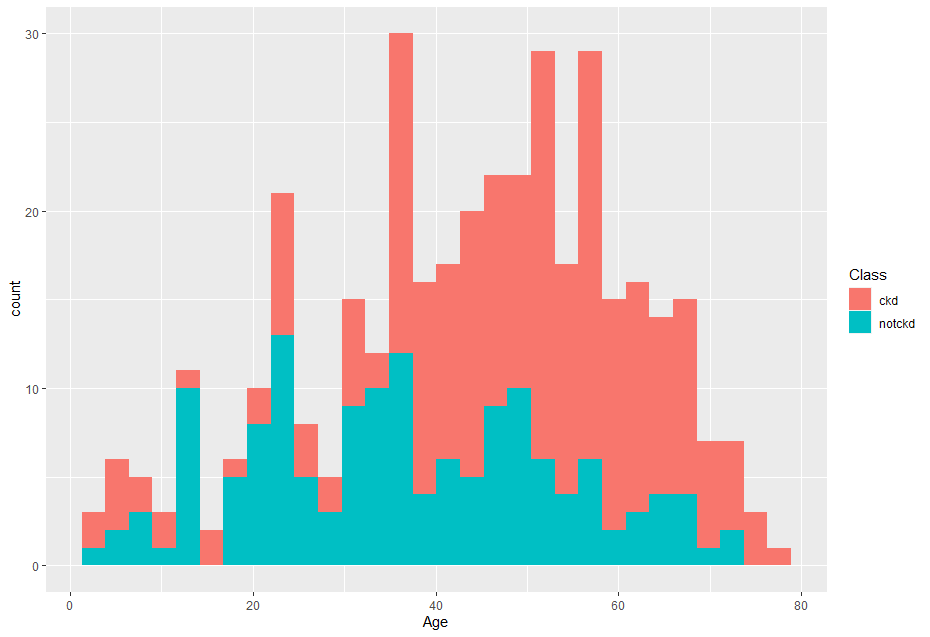
****

HH

**Histogram**

Count of People in the range of age with CKD and NOTCKD

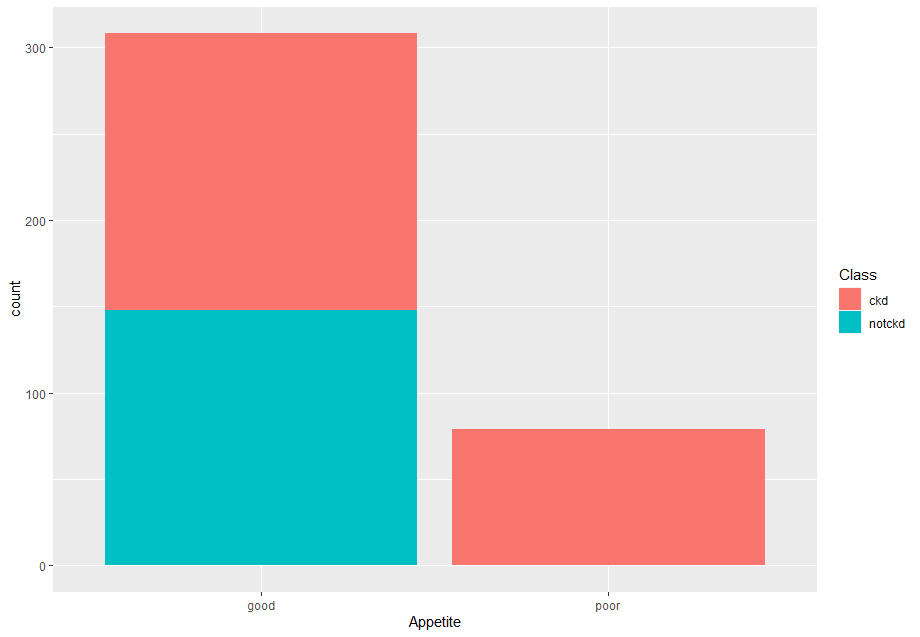


****

**Bar Graph**

People with good and bad appetite compared to having CKD or Not

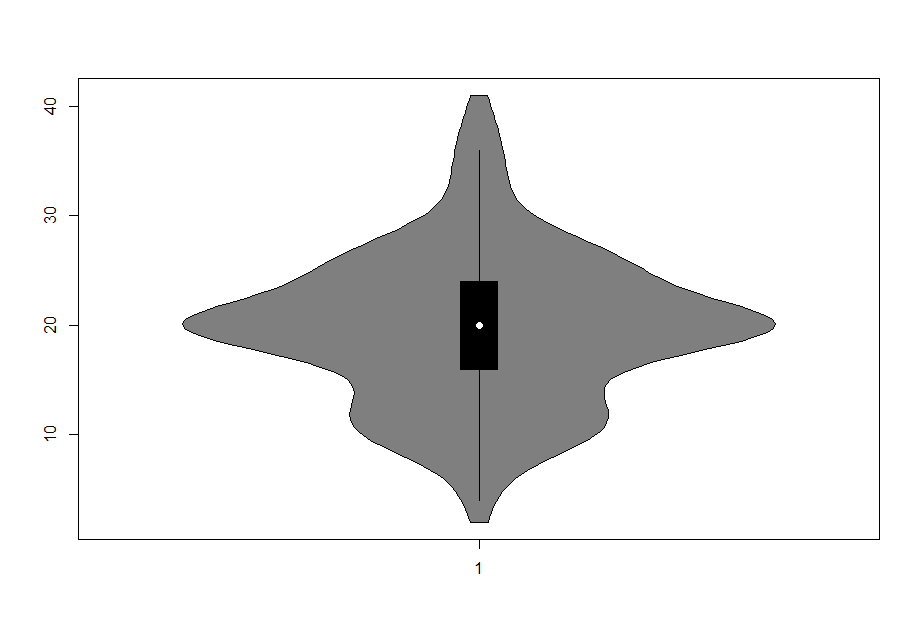


****

**Violin Graph**

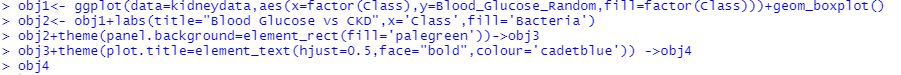
This graph represents the spread of potassium level

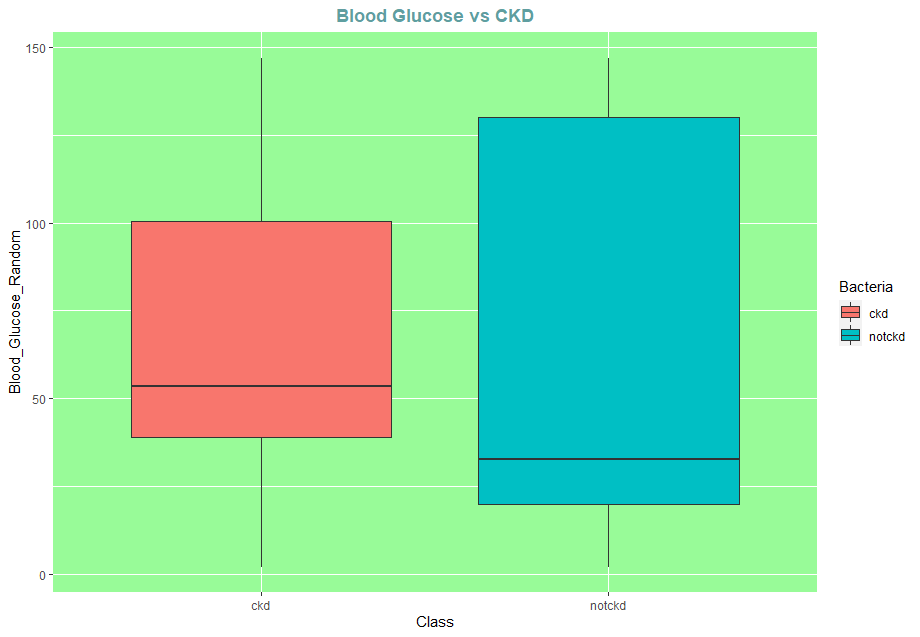


****

**THEME BASED GRAPH**

Comparison with Blood Glucose Levels and CKD



****

**ML ALGORITHM**

Since it’s a binary classification problem, we should use Classification Based ML algorithms like Logistic Regression or Decision Tree.

For this dataset we have implemented Logistic Regression and Decision Tree Algorithms

**MODEL CREATION**

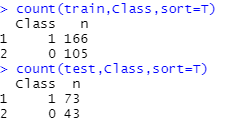
First we will be changing target column to numeric since it’s a binary classification we can assign 0 and 1

****

We need to split the data into training set and testing test, here we are splitting it in 70:30 raito. The splitting is also randomised.



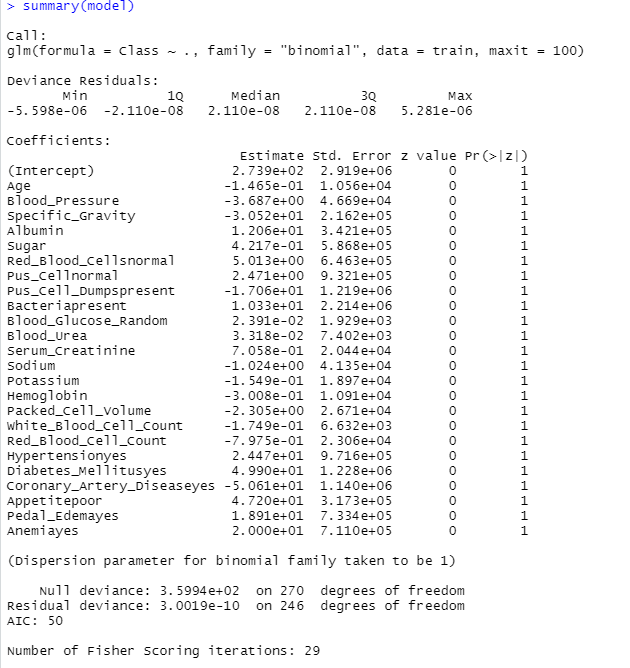
Count of train and test dataset



**Logistic Regression**

Fitting the train dataset into General Linear Model function using binomial family and dataset’s target column along with all other column, calling the summary function to analyse our models statistic values

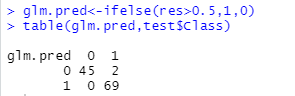
****

****

Predicting the values using test dataset



Creating the confusion matrix



Calculating the accuracy

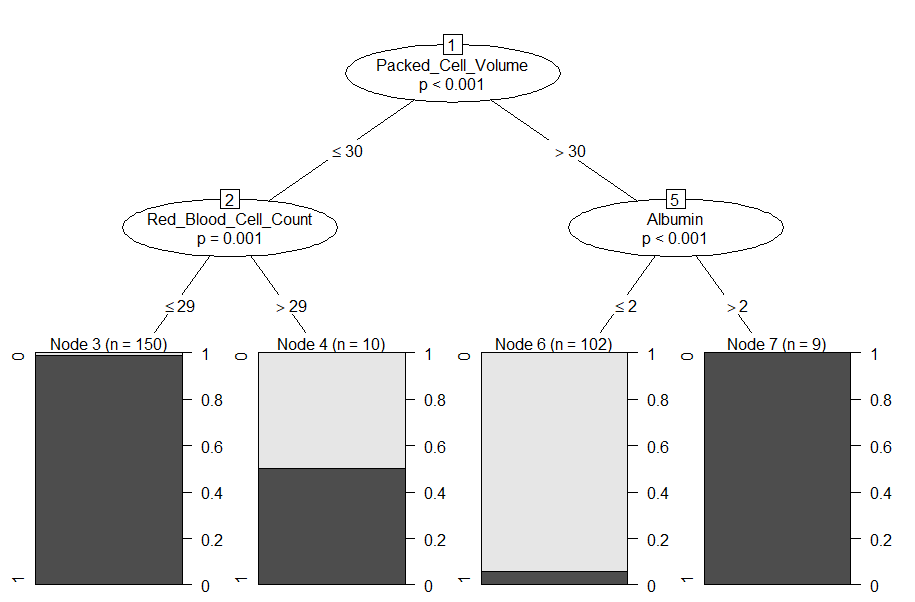


We can see approximate to 98% of accuracy.

**Decision Tree Algorithm**

We will be using ctree function for fitting decision tree model. Also make sure all columns are of type numeric or factor.



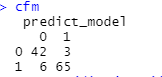


Predicting using our model



Creating the confusion matrix





Calculating the accuracy





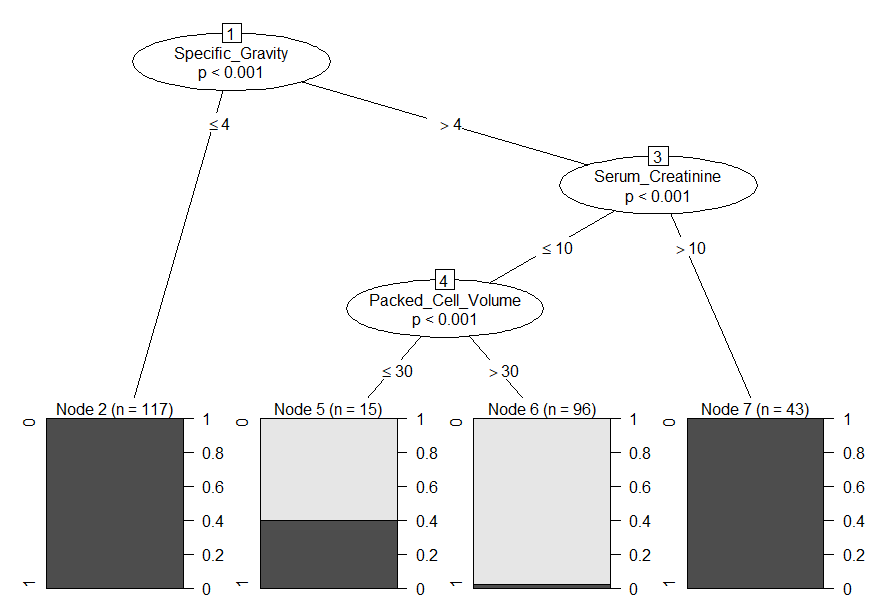
Thus the accuracy is 86% when using all the columns

**Decision tree with numeric column**

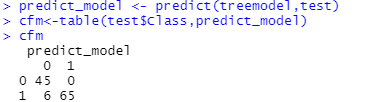
We are training our model only with numeric column in order to improve the accuracy of the model







Predicting and creating the confusion matrix



Calculating the accuracy

We get an accuracy of 94% which is better compared to our model which had 86%



**RESULT**

**The Chronic Kidney Disease dataset has been used to predict whether the person would suffer from Chronic Kidney Disease or not using logistic regression algorithm and Decision Tree Algorithm and we find that Logistic Regression predicts better compared to Decision Tree. Here all columns were used for fitting inside the model. The accuracy may not accurate due to less number of data, hence more data should be trained and test in order to perform accurately in real world.**