PREDICTIVE ANALYSIS ON CHRONIC KIDNEY DISEASE

Project Report submitted to the

SDM COLLEGE (Autonomous)

in partial fulfilment of the degree of

MASTER OF SCIENCE

IN

STATISTICS

by

Samskruthi H D

Under the supervision of

Asst. Prof. Ms. Supriya Shivadasan Padmavati

Department of Post Graduate Studies in Statistics

SRI DHARMASTHALA MANJUNATHESHWARA

COLLEGE (Autonomous)

UJIRE - 574240

Karnataka, INDIA

AUGUST 2022

SRI DHARMASTHALA MANJUNATHESHWARA COLLEGE
(AUTONOMOUS)
UJIRE - 574240

DEPARTMENT OF STATISTICS

CERTIFICATE

Certified that this is the bonafide record of project work done by Ms. Samskruthi H D during the year 2022 as a part of her M.Sc (Statistics) fourth semester course work.

Reg. No.	**2**	**0**	**1**	**9**	**4**	**2**

Project Guide **Head of the Department**

Examiner

1.

2.

Date:
Place: Ujire

DECLARATION

I, Samskruthi H D, hereby declare that the matter embodied in this report entitled '**Predictive Analysis on Chronic Kidney Disease**' is a bonafide record of project work carried out by me under the guidance and supervision of **Asst. Prof. Ms. Supriya Shivadasan Padmavati**, Department of Statistics, SDM College, Ujire - 574240, Karnataka, India. I further declare that no part of the work contained in the report has previously been formed the basis for the award of any Degree, Diploma, Associateship, Fellowship or any other similar title or recognition of any other university.

Date:
Place: Ujire

(SAMSKRUTHI H D)
E-mail:samskruthihd00@gamil.com

CERTIFICATE

This is to certify that the project report entitled '**Predictive Analysis on Chronic Kidney Disease**' is a bonafide record of an authentic work carried out by **Samskruthi H D**, under my guidance and supervision in the Department of Post Graduate Studies and Research in Statistics, SDM College, Ujire, in partial fulfilment of the requirements for the award of the degree of Master of Science in Statistics, under Mangalore University, Mangalagangothri. I further certify that this report or part thereof has not previously been presented or submitted elsewhere for the award of any Degree, Diploma, Associateship, Fellowhsip or any other similar title of any other institution or university.

Date:
Place: Ujire

(Supriya Shivadasan Padmavati)
E-mail: supriyasp@sdmcujire.in

ACKNOWLEDGEMENTS

Firstly, I would like to thank our Principal ***Dr. P. N. Udayachandra*** *for providing the necessary facilities for the completion of this project work in our college.*

I would also thank our Dean ***Dr. Vishwanath P*** *for his support.*

It is my privilege to thank our HOD ***Prof. Shanthiprakash*** *for his suggestions and support.*

I am very grateful to my Research Supervisor, ***Asst. Prof. Ms. Supriya Shivadasan Padmavati****, Department of Statistics, SDM College, Ujire, for his kind help and encouragement throughout my project work.*

I gratefully acknowledge my teachers at the Department of Statistics, SDM College, Ujire, ***Asst. Prof. Ms. Shwetha Kumari*** *and* ***Asst. Prof. Mr. Pradeep K*** *for their support during my project work.*

I am also thankful to all my family members and friends for their constant encouragement and help in each step.

My sincere thanks also goes to the students of SDM College, Ujire, who have helped me directly or indirectly during my project work.

*Finally to all who helped me in many ways, I say, '****Thank You****!'.*

(Samskruthi H D)

Contents

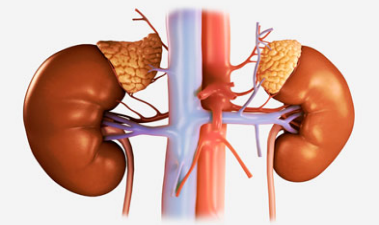
List of Tables

List of Figures

1 Chapter 1

1.1 Introduction

Chronic Kidney Disease



The disability of the kidneys to perform their regular blood filtering function is called Chronic Kidney Disease (CKD). The term chronic describes the slow degradation of the kidney cells over a long period of time. Chronic kidney disease is a condition in which the kidneys are damaged and cannot filter blood as well as they should. Because of this, excess fluid and waste from blood remain in the body and may cause other health problems, such as heart disease, stroke, increased occurances of infections, Depression or lower quality of life, low number of red blood cells. CKD has varying levels of seriousness. It usually gets worse over time though treatment has been shown to slow progression. If left untreated, CKD can progress to kidney failure and early cardiovascular disease. When the kidneys stop working, dialysis or kidney transplant is needed for survival.

Initially there are generally no symptoms; later, symptoms may include leg swelling, feeling tired, vomiting, loss of appetite, and confusion. Complications can relate to hormonal dysfunction of the kidneys and include high blood pressure, bone disease, and anemia. Additionally CKD patients have markedly increased cardiovascular complications with increased risks of death and hospitalization.

There are a few other conditions or circumstances that can cause kidney disease,

1. **Glomerulonephritis**: Glomerulonephritis is a group of diseases that cause inflammation and damage the kidney's filtering units. These disorders are the third most common type of kidney disease.

2. **Inherited diseases**: Polycystic kidney disease, or PKD, is a common inherited disease that causes large cysts to form in the kidneys and damage the surrounding tissue.

3. **Kidney and urinary tract abnormalities before birth**: Malformations that occur as a baby develops in its mother's womb. For example, a narrowing may occur that prevents normal outflow of urine and causes urine to flow back up to the kidney. This causes infections and may damage the kidneys.

4. **Autoimmune diseases**: When the body's defense system, the immune system, turns against the body, it's called an autoimmune disease. Lupus nephritis is one such autoimmune disease that results in inflammation (swelling or scarring) of the small blood vessels that filter wastes in your kidney.

5. **Other causes**: Obstructions caused by kidney stones or tumors can cause kidney damage. An enlarged prostate gland in men or repeated urinary infections can also cause kidney damage.

Signs and symptoms of chronic kidney disease develop over time if kidney damage progresses slowly. Loss of kidney function can cause a buildup of fluid or body waste or electrolyte problems. Depending on how severe it is, loss of kidney function can cause:

1. Nausea
2. Vomiting
3. Loss of appetite
4. Fatigue and weakness
5. Sleep problems
6. Urinating more or less
7. Decreased mental sharpness
8. Muscle cramps
9. Swelling of feet and ankles
10. Dry, itchy skin

11. High blood pressure (hypertension) that's difficult to control
12. Shortness of breath, if fluid builds up in the lungs
13. Chest pain, if fluid builds up around the lining of the heart

Signs and symptoms of kidney disease are often nonspecific. This means they can also be caused by other illnesses. Because your kidneys are able to make up for lost function, you might not develop signs and symptoms until irreversible damage has occurred.

Chronic kidney disease occurs when a disease or condition impairs kidney function, causing kidney damage to worsen over several months or years. Diseases and conditions that cause chronic kidney disease are Type 1 or type 2 diabetes, High blood pressure, Polycystic kidney disease or other inherited kidney diseases, Prolonged obstruction of the urinary tract, from conditions such as enlarged prostate, kidney stones and some cancers, Recurrent kidney infection. Factors that can increases the risk of chronic kidney disease are Diabetes, High blood pressure, Heart (cardiovascular) disease, Smoking, Family history of kidney disease, Abnormal kidney structure, Older age, Frequent use of medications that can damage the kidneys.

Chronic kidney disease can affect almost every part of your body. Potential complications include:

- Fluid retention, which could lead to swelling in your arms and legs, high blood pressure, or fluid in your lungs (pulmonary edema)
- A sudden rise in potassium levels in your blood (hyperkalemia), which could impair your heart's function and can be life-threatening
- Anemia
- Heart disease
- Weak bones and an increased risk of bone fractures
- Decreased sex drive, erectile dysfunction or reduced fertility
- Damage to your central nervous system, which can cause difficulty concentrating, personality changes or seizures

- Decreased immune response, which makes you more vulnerable to infection
- Pericarditis, an inflammation of the saclike membrane that envelops your heart (pericardium)
- Pregnancy complications that carry risks for the mother and the developing fetus
- Irreversible damage to your kidneys (end-stage kidney disease), eventually requiring either dialysis or a kidney transplant for survival

Kidney Disease Treatment- Some forms of kidney disease are treatable. The goals of these treatments are to ease symptoms, help keep the disease from getting worse, and lessen complications. In some cases, your treatment may help restore some of your kidney function. There is no cure for chronic kidney disease. The plan you and your doctor will decide on will depend on what's causing your kidney disease. In some cases, even when the cause of your condition is controlled, your kidney disease will worsen.

Once your kidneys can't keep up with waste on their own, you'll have treatment for end-stage kidney disease. This can include:

Dialysis: Waste and extra fluid are taken out of your body when your kidneys can't do it anymore. There are two types:

1. Hemodialysis, where a machine removes the waste and extra fluids from your blood. In hemodialysis, the blood is pumped through a special machine that filters out waste products and fluid. Hemodialysis is done at your home or in a hospital or dialysis center. Most people have three sessions per week, with each session lasting 3 to 5 hours. However, hemodialysis can also be done in shorter, more frequent sessions. Several weeks before starting hemodialysis,

   most people will have surgery to create an arteriovenous (AV) fistula. An AV fistula is created by connecting an artery and a vein just below the skin, typically in the forearm. The larger blood vessel allows an increased amount of blood to flow continuously through the body during hemodialysis treatment. This means more blood can be filtered and purified. An arteriovenous graft (a looped, plastic tube) may be implanted and used for the same purpose if an artery and vein cannot be joined together. The most common side effects of hemodialysis are low blood pressure, muscle cramping, and itching.

2. Peritoneal dialysis, which involves inserting a thin tube called a catheter into your abdomen. Then, a solution goes into your abdomen that absorbs the waste and fluids. After a while, the solution drains from your body. In peritoneal dialysis, the peritoneum (membrane that lines the abdominal wall) stands in for the kidneys. A tube is implanted and used to fill the abdomen with a fluid called dialysate. Waste products in the blood flow from the peritoneum into the dialysate. The dialysate is then drained from the abdomen. There are two forms of peritoneal dialysis, continuous ambulatory peritoneal dialysis, where the abdomen is filled and drained several times during the day, and continuous cycler-assisted peritoneal dialysis, which uses a machine to cycle the fluid in and out of the abdomen at night while the person sleeps. The most common side effects of peritoneal dialysis are infections in the abdominal cavity or in the area where the tube was implanted. Other side effects may include weight gain and hernias. A hernia is when the intestine pushes through a weak spot or tear in the lower abdominal wall.

Kidney transplant: A surgeon replaces your kidney with a healthy one from a donor. This donor can be living or deceased. After the procedure, you take medicine for the rest of your life to make sure that your body doesn't reject your new kidney.

1.2 Motivation

Chronic kidney disease (CKD) is a global public health problem. It affects 50 million people worldwide, among them, over 1.7 million have end-stage renal disease (ESRD), requiring either dialysis or transplantation. In 2004, about half a million Americans had kidney failure. This number is projected to reach 1.8 million by 2020. When comparing the United States with other countries, the United States has one of the highest incidences of ESRD, ranking below Mexico only.
It is expensive to take care of patients with CKD. It cost Medicare nearly 25 billion dollars to care for individuals who are on dialysis or received kidney transplantation. According to USRDS 2010 data, CKD comprised 6.8% of Medicare population, but used over 14% of Medicare expenditures. CKD is a silent disease, meaning an individual who has this disease does not usually experience any symptoms, even when the individual has reached a late stage of the condition such as requiring dialysis. In addition, the majority of people who have CKD do not know they have the disease. Kidney disease is more common in Asian-Americans, African-Americans, and Hispanics.

Chronic kidney disease is an irreversible and silent condition; you may not know you have it until the disease has reached a very late stage. This project will provide detection of risk factors of CKD and result in CKD prevention.

1.3 Literature Review

In 2021 Gazi Mohammed Ifraz carried out a Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods. The main motivation of this study is to predict renal disease by analyzing data from those indices and applying three machine learning classification approaches to predict the disease, then choosing the approach with the highest accuracy rate. Two classification techniques are used:-nearest neighbors classifier, decision tree classifier (DT). According to the findings of the study, the decision tree approach can be used to predict chronic kidney disease more accurately. According to the study, their precision was 96.25 percent, and their accuracy was 97 percent.

In 2019, October S. Revathy and M Ramesh carried out a Analysis on Chronic kidney disease dataset. The main objective of this study is to predict chronic kidney disease in patients. The techniques used to predict CKD using the classifiers like Decision Tree, Random Forest and Support Vector Machine and also suggested best prediction model. The performance of the models are evaluated based on the accuracy of prediction. The results of the research showed that Random Forest Classifier model better predicts CKD in comparison to Decision trees and Support Vector machines. The comparision can also be done based on the time of execution, feature set selection as the improvisation of this research.

In 2021, Surya Krishnamurthy carried out a analysis on Chronic kidney disease Disease Using National Health Insurance Claim Data in Taiwan. The main aim is to develop machine-learning models that predict the onset of CKD. The various modeling algorithms used to perform the analysis are deep neural networks (CNN and BLSTM), LightGBM, decision tree. The results of this study is concluded as follows: The above models predict patients risk of developing chronic kidney disease after a period of 6 or 12 months. Among various models tested, convolutional neural networks (CNN) performed best, with an AUROC metric of 0.957 and 0.954 for 6 and 12 months, respectively.

In 2022 Jan 28, Aman Preet Gulati carried out a analysis based on Chronic kidney disease dataset. The main goal of this study is to predict whether an individual will have chronic kidney disease or not based on the data. The machine learning

techniques used to predict Chronic kidney Disease are data processing, Exploratory data analysis and model building using the models such as logistic regression and k-nearest neighbour. The result of this study is concluded as follows: Among various models Logistic Regression performs better to predict Chronic kidney disease patients with accuracy 0.97.

In 2021 january, Pankaj Chittora carried out a research based on Prediction of Chronic KIdney Disease. The main objective of this study is to predict Chronic Kidney Disease based on full features and important features of CKD dataset. For feature selection three different techniques have been applied: correlation-based feature selection, Wrapper method and LASSO regression. In this perception, five classifiers algorithm were applied viz. KNN, Artificial neural network, CHAID, linear support vector machine(LSVM), and random tree. It was observed that LSVM achieved the highest accuracyof 98.86% in Synthetic Minority Oversampling Technique (SMOTE) with full features. In this research, KNN did not give suitable result. As per the result, it is concluded that SMOTE is a best technique for balancing a dataset. LSVM achieved the highest accuracy in all experiments as compared to other classifiers.

In 2022 january, Domor I Mienye conducted a analysis on Prediction of Chronic Kidney Disease Using Feature Selection and Boosted Classifiers. The least informative features are discarded, and only the relevant features were used in building the CKD prediction model. Secondly, six machine learning models were developed using the LR, SVM, DT, AdaBoost-LR, AdaBoost-SVM, and AdaBoost-DT algorithms. The models were trained using both the complete feature set and the reduced feature set. Meanwhile, the boosted decision tree (AdaBoost-DT) achieved the best performance with a value of 1.000 in all the four performance evaluation metrics, i.e.accuracy, precision, sensitivity, and F-measure.

In 2015, Vijarani Mohan carried out a research on kidney disease prediction using SVM and ANN algorithms. In this research work classification process is used to classify four types of kidney diseases. Comparisons of Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms are done based on the performance factors classification accuracy and execution time. From the results, it can be concluded that the ANN achieves increased classification performance, yields results that are accurate, hence it is considered as best classifier when compared with SVM classifier algorithm.

In 2020 Koushal Kumar conducted a analysis on Kidney Disease. In this analysis they used Learning vector quantization (LVQ), two layers feed forward perceptron

trained with back propagation training algorithm and Radial basis function (RBF) networks for diagnosis of kidney stone disease. They have compared the performance of all three neural networks on the basis of its accuracy, time taken to build model, and training data set size. They used Waikato Environment for Knowledge Analysis (WEKA) tool for execution. Finally from the experimental results, authors concluded that multilayer perceptron trained with back propagation is best algorithm for kidney stone diagnosis.

In 2016, Jyothi Saini carried out a research on Design and Analysis of algorithm for prediction of Kidney disease. The main objective is to predict kidney disease using data mining tools such as Support Vector Machine(SVM) and Artificial Neural Network(ANN). These two techniques are best in their own ways. The performance of the models are evaluated based on the accuracy of prediction. According to findings of their study, it is found that there is still a lot of future work need to be done. In future these Clusters can more improved by using techniques.

1.4 Objectives

1. To study about the age of kidney disease patients.
2. To study about the factors affecting the kidney disease in patients.
3. To check the association between appetite and presence of kidney disease.
4. To check whether the hypertension and kidney disease are dependent.
5. To determine the factors that causes the kidney disease in patients.
6. To predict the presence of Chronic Kidney Disease in patients.

1.5 Scope of the study

The results of this project can be used as a source of basic information regarding kidney disease, which will helpful for the researchers, doctors, etc., Also it will helpful for the Health Departments. And also these results will be useful for those who are interested to discover new methods or treatments to prevent kidney disease. And this is helpful in finding the risk factors affecting the kidney disease in patients.

2 Chapter 2

2.1 Materials and Methods

About the data
A secondary data has been collected from the website of UCI Machine Repository(Chronic Kidney Disease). This data is collected in the year 2015. The data contains 400 observations and 21 variables. Each variable corresponds to an individual kidney disease patients. The description about the variables considered in the analysis are given as follows:

- **id**-Id of each patient.
- **age**-Age of an individual patient.
- **bp**-Blood pressure of the kidney disease patients.
- **sg**-Specific gravity of the urine. The higher sg indicates dehydration.
- **al**-Albumin level, it is a type of protein present in blood.
- **su**-sugar level present in the blood
- **pcc**-Pus cell clumps present or not. It is a blood cell present in the urine, it causes some of the infections.
- **ba**-Bacteria present or not.
- **bu**-Blood urea, it is a serious condition occurs when kidneys are damaged, it is a nitrogenous end products of metabolism.
- **sc**-Serum creatinine,it is chemical compound left from energy processes in muscles.
- **sod**-Sodium content present in the body.
- **pot**-Potassium content present in the body.
- **hemo**-Hemoglobin level present in the blood.
- **rc**-Red blood cells count present in the blood.
- **htn**-Patient having hypertension or not(yes/no).

- **dm**-Patient having Diabetes or not(yes/no).
- **cad**-Patient having coronary artery disease or not(yes/no)
- **appet**-Appetite is a patients desire to eat food. Whether it is good or poor.
- **pe**-Patient having pedel edema disease or not(yes/no).
- **ane**-Patient having anemia disease or not(yes/no).
- **classification**- Patient having kidney disease or not. Categorized as ckd and notckd.

2.2 Statistical Techniques Used for Analysis

Tools Used:

The open source softwares and programming languages 'R' and 'Python' has been used to carry out the analysis of the data. The statistical methods considered in order to carry out the analysis are given as follows:

2.2.1 Fisher's Exact Test:

Fisher's exact test is used to determine whether or not there is a significant association between two categorical variables. It is typically used as an alternative to the Chi-Square Test of Independence when one or more of the cell counts in a 2x2 table is less than 5. As an exact significance test, Fisher's test meets all the assumptions on which basis the distribution of the test statistic is defined. In practice, this means that the false rejection rate equals the significance level of the test, which is not necessarily true for approximate tests such as χ^2 the test.

There are certain assumptions on which the Fisher Exact test is based.

1. It is assumed that the sample that has been drawn from the population is done by the process of random sampling. This assumption is also assumed in general in all the significance tests.
2. In the Fisher Exact test, a directional hypothesis is assumed. The directional hypothesis assumed is nothing but the hypothesis based on the one tailed test. In other words, the directional hypothesis assumed is that type of hypothesis which predicts either a positive association or a negative association, but not both.

3. In the Fisher Exact test, mutual exclusivity within the observations is assumed. In other words, the given case should fall in only one cell in the table. The dichotomous level of measurement of the variables is assumed.

The hypothesis under consideration are,
H_0(Null hypothesis): There is no relationship between the variables.

H_1(Alternative hypothesis): There is relation between the variables.

The test statistic is given as follows,

$$p = \frac{\binom{n_{1,1}+n_{1,2}}{n_{1,1}}\binom{n_{2,1}+n_{2,2}}{n_{2,1}}}{\binom{n_{1,1}+n_{1,2}+n_{2,1}+n_{2,2}}{n_{1,1}+n_{2,1}}}$$

The test procedure is to reject the null hypothesis H_0 if the $p-value$ is less than the alpha value, then reject the null hypothesis H_0 and accept the alternative hypothesis H_1. Generally alpha value of 0.05 is chosen.This alpha value denotes the probability of erroneously rejecting null hypothesis when it is true.

2.2.2 Mann Whitney U Test:

The modules on hypothesis testing presented techniques for testing the equality of means in two independent samples. An underlying assumption for appropriate use of the tests described was that the continuous outcome was approximately normally distributed or that the samples were sufficiently large (usually n1¿ 30 and n2¿ 30) to justify their use based on the Central Limit Theorem. When comparing two independent samples when the outcome is not normally distributed and the samples are small, a nonparametric test is appropriate. A popular nonparametric test to compare outcomes between two independent groups is the Mann Whitney U test. The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape). Some investigators interpret this test as comparing the medians between the two populations.

The null and alternative hypotheses for the nonparametric test are stated as follows:

H_0(Null hypothesis): The medians of all groups are same

H_1(Alternative hypothesis): At least one group median is not equal to the others.

The test procedure is to reject the null hypothesis H_0 if the $p - value$ is less than the level of significance($\alpha = 0.05$), then reject the null hypothesis H_0 and accept the alternative hypothesis H_1.

2.2.3 Two sample Z-test for Proportions:

In statistics, a two-sample z-test for proportions is a method used to determine whether two samples are drawn from the same population. This test is used when the population proportion is unknown and there is not enough information to use the chi-squared distribution. The test uses the standard normal distribution to calculate the test statistic. While performing the test, Z-statistic is computed from two independent samples.

The hypothesis of this test are as follows,

H_0(Null hypothesis): The two proportions are equal.

H_1(Alternative hypothesis): The two proportions are not equal.

In order to be able to use the two-sample z-test, the following conditions must be met:

- The two populations must be normal or approximately normal
- The two samples must be randomly sampled from the two populations
- The two proportions must be independent

If any of the above conditions are not met, the two-sample z-test cannot be used and another test must be selected. The two-sample z-test is advantageous because it does not require any knowledge of the population standard deviation.

There are two steps in conducting a two-sample z-test for proportions.

- The first step is to calculate the standard error of the difference between the two population proportions.

- The second step is to calculate the z-test statistic. This is done by taking the difference between the two population proportions and dividing it by the standard error of the difference.

Once the z-test statistic is calculated, the Z-table can be used to determine whether the two population proportions are different. If the z-statistic is greater than or equal to the critical value or level of significance, then it can be concluded that there is enough evidence that there exists a difference between the two population proportions. And, the null hypothesis can thus be rejected.

2.2.4 Logistic Regression:

Logistic regression is a statistical model to predict relationship between categorical dependent variable(target variable) and one or more independent variable(continous predictor variable).

It is a special case of generalized linear model where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic regression predicts the probability of occurance of binary event utilizing a logit function.

Linear regression equation is given as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

Where Y is dependent variable and X_1, X_2.....and X_n are exploratory variables.

Sigmoid function is given by,

$$P = 1/1 + e^{-y}$$

Apply sigmoid function on linear regression,

$$P = 1/1 + e^{-(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)}$$

The dependent variable in logistic regression follows bernoulli distribution. Logistic regression is estimated using Maximum Likelihood Estimation(MLE) approach.

2.2.5 Countplot:

Seaborn is a module in Python that is built on top of matplotlib that is designed for statistical plotting. Seaborn can create all types of statistical plotting graphs. One of the plots that seaborn can create is a countplot. A countplot is kind of like a histogram or a bar graph for some categorical area. It simply shows the number of occurrences of an item based on a certain type of category.

2.2.6 Decision Tree Classifier:

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Types of decision trees are based on the type of target variable we have. It can be of two types:

- Categorical Variable Decision Tree: Decision Tree which has a categorical target variable then it called a Categorical variable decision tree.
- Continuous Variable Decision Tree: Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree.

The common problem with Decision trees, especially having a table full of columns, they fit a lot. Sometimes it looks like the tree memorized the training data set. If there is no limit set on a decision tree, it will give you 100% accuracy on the training data set because in the worse case it will end up making 1 leaf for each observation. Thus this affects the accuracy when predicting samples that are not part of the training set. Pruning Decision Tree is used to remove overfitting.

Pruning Decision Trees
The splitting process results in fully grown trees until the stopping criteria are reached. But, the fully grown tree is likely to overfit the data, leading to poor accuracy on unseen data. In pruning, you trim off the branches of the tree, i.e.,

remove the decision nodes starting from the leaf node such that the overall accuracy is not disturbed. This is done by segregating the actual training set into two sets: training data set, D and validation data set, V. Prepare the decision tree using the segregated training data set, D. Then continue trimming the tree accordingly to optimize the accuracy of the validation data set, V.

2.2.7 Random forest classifier:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which we'll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged, and for a classification task, a majority vote i.e. the most frequent categorical variable will yield the predicted class.

3 Chapter 3

Results and discussion

3.1 Characteristics of Sample Respondents

3.1.1 Pus cell clumps and Bacteria

The following table shows the frequency of the patients having pus cell clumps and bacteria considered in this study.

Table 1: Frequency table for pus cell clumps and bacteria

	pcc	ba
present	46 (11.5%)	26 (6.5%)
notpresent	354 (88.5%)	374 (93.5%)

Here we can observe that, pus cell clumps is present in 11.5% of the patients and bacteria is present in 6.5% of the patients considered in this study.

3.1.2 Hypertension, Diabetes mellitus, Coronary artery disease, Pedel edema and Anemia

The following table shows the frequency of the patients having hypertension, diabetes mellitus, coronary artery disease, pedel edema and Anemia considered in this study.

Table 2: Frequency table of symptoms based on class of kidney disease

	yes	no
htn	147 (36.75%)	253 (63.25%)
dm	137 (34.25%)	263 (65.75%)
cad	34 (8.5%)	366 (91.5%)
pe	76 (19%)	324 (81%)
ane	60 (15%)	340 (85%)

Here we can observe that, 36.75% patients are having hypertension, 34.25% patients are having diabetes mellitus, 8.5% patients are having coronary artery disease,

19% patients are having pedel edema and 15% patients are having anemia considered in this study.

3.1.3 Appetite

The following table shows the frequency of the patients having appetite considered in this study

Table 3: Frequency table for appetite disease

appet	Frequency
good	317 (79.25%)
poor	83 (20.75%)

From the table we can observe that, appetite is good in 79.25% of the patients and poor in 20.75% of the patients considered in this study.

3.1.4 Class

The following table shows the frequency of the patients having kidney disease considered in this study

Table 4: Frequency table of class of kidney disease

Class	Frequency
ckd	250 (62.5%)
notckd	150 (37.5%)

From the table we can observe that, 62.5% of the patients are having kidney disease and 37.5% of the patients are not having kidney disease considered in this study.

3.2 Univariate analysis of Factor Age

3.2.1 To study about the age of kidney disease patients.

The following table shows the frequency of the age of the kidney disease patients.

Table 5: Frequency table of class of kidney disease based on age group

Age group	kidney disease	
	yes	no
0-14(childrens)	13 (3.25%)	1 (0.25%)
15-24(youths)	7 (1.75%)	13 (3.25%)
25-64(adults)	153 (38.25%)	115 (28.75%)
65-100(elders)	77 (19.25%)	21 (5.25%)

The below plot shows age of the kidney disease patients.

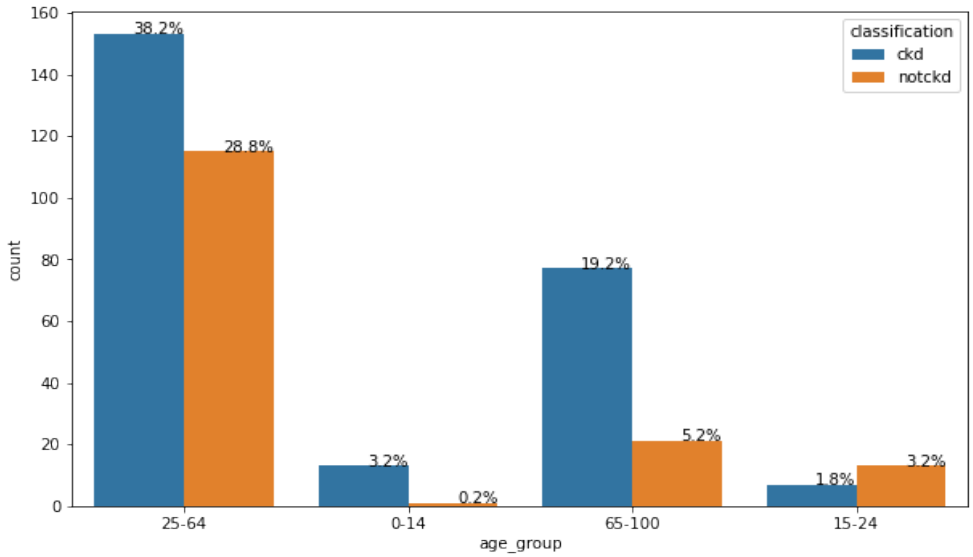


Figure 1: Count plot for age group of patients

Here we can observe that, 38.25% of the respondents considered in this study are having kidney disease between the age group 25-64. In age group between 65-100,

19.25% of the patients having kidney disease. 3.25% of the patients having kidney disease between the age group 0-14. And 1.75% of the patients having kidney disease between the age group 15-24. Hence, we can say that adults are having highest chance of getting kidney disease than the elders.

The average age of respondents at which kidney disease occurred is 51.48 and the standard deviation is 16.97.

3.3 Bivariate analysis using Fisher's Exact test of Independence.

3.3.1 Testing the association between appetite and class of the Kidney Disease

The following table shows patients with kidney disease having good or poor appetite and patients with no kidney disease having good or poor appetite.

Table 6: Contingency table on class and appetite symptom

Class Appetite	ckd	notckd
good	168	149
poor	82	1

The Hypothesis are as follows:

H_0 : The variables Appetite and Classification are independent.

H_1 : The variables Appetite and Classification are dependent.

The obtained values are as follows:

- p-value = 2.2e-16
- Cramer V = 0.3836

Here we can observe that the obtained p-value is less than 0.05. Hence we do reject the null hypothesis. And we conclude that the variables Appetite and Classification are dependent.

Therefore, there is some relationship between patients with good or poor appetite and classification of the kidney disease patients. From the Cramer V value, we conclude that the strength of the association between appetite and classification is 0.3836. Hence, there is a weak association between appetite and classification.

3.3.2 Testing the association between hypertension and class of of the Kidney Disease

The following table shows the classification of the kidney disease patients and hypertension of the respondents.

Table 7: Contingency table on class and hypertension disease

Class Hypertension	ckd	notckd
no	103	150
yes	147	0

The Hypothesis are as follows:

H_0 : The variables Hypertension and Classification are independent.

H_1 : The variables Hypertension and Classification are dependent.

The obtained values are as follows:

- p-value = 1.99e-16
- Cramer V = 0.5904

Here we can observe that the obtained p-value is less than 0.05. Hence we do reject the null hypothesis. And we conclude that the variables Hypertension and Classification are dependent.

Therefore, there is some relationship between hypertension of the patients sand classification of the kidney disease patients. From the Cramer V value, we conclude that the strength of the association between hypertension and classification is 0.5904. Hence, we conclude that there is a strong association between hypertension and classification.

Conclusion:

1. From the univariate analysis of factor age, we observe that adults are having highest chance of getting kidney disease compare to all other groups. Therefore, the average age of people at which kidney disease occurred is approximately 51.

2. From the bivariate analysis using fishers exact test of independence, we observe that there is a weak association between appetite and classification of the kidney disease patients. Also, the strength of the association between appetite and classification is 0.3836.

3. From the bivariate analysis using fishers exact test of independence, we observe that there is a strong association between hypertension and classification of the kidney disease patients. Also, the strength of the association between hypertension and classification is 0.5904.

3.4 To study about the factors affecting the kidney disease in patients.

3.4.1 Analysis of factors affecting the kidney disease using Mann-Whitney U test

To check normality assumption, Anderson-Darling normality test is applied. The hypothesis are as follows:

H_0: The data follows normal distribution

H_1: The data does not follow normal distribution

The below table shows the results of Anderson darling normality test.

Table 8: Results of Anderson Darling normality test

Factors	**Test statistic (A)**		**p-value**		**Decision**
	ckd	notckd	ckd	notckd	
blood pressure	9.1739	14.164	<2.2e-16	2.2e-16	reject H_0
specific gravity	12.122	23.301	1.2e-16	<2.2e-16	reject H_0
albumin	10.208	55.582	2.01e-10	0.0112	reject H_0
sugar	33.607	51.582	0.0002	1.2e-14	reject H_0
blood urea	4.9135	2.918	3.45e-12	2.28e-07	reject H_0
seram creatinine	35.202	10.099	1.98e-11	0.0233	reject H_0
sodium	10.855	2.4028	2.2e-16	4.17e-06	reject H_0
potassium	12.524	7.554	1.54e-12	2.52e-12	reject H_0
hemoglobin	4.1327	0.6966	2.68e-10	0.0675	reject H_0
red blood cell	37.985	5.5337	<2.2e-16	1.05e-13	reject H_0

The assumption of normality does not met. The p-value of blood pressure, specific gravity, albumin, sugar, blood urea, seram creatinine, sodium, potassium, hemoglobin and red blood cell is less than 0.05. Hence we do reject the null hypothesis. Therefore we can conclude that data does not follow normal distribution.

Hence we use non parametric alternative, Wilcoxon Mann-Whitney U test to carry-out the analysis.

The hypothesis of Mann-Whitney U test are as follows:

H_0: The median of all groups are same
H_1: At least one group median is not equal to the others

The following table shows the results of Mann-Whitney U test for presence of kidney disease and factors affecting the kidney disease in patients.

Table 9: Results of Mann Whitney U-test

Factors	**W-Statistic**	**p-value**	**Decision**
blood pressure	160000	2.2e-16	reject H_0
specific gravity	159900	<2.2e-16	reject H_0
albumin	89825	0.001242	reject H_0
sugar	39875	<2.2e-16	reject H_0
blood urea	160000	<2.2e-16	reject H_0
seram creatinine	131375	<2.2e-16	reject H_0
sodium	160000	2.2e-16	reject H_0
potassium	160000	<2.2e-16	reject H_0
hemoglobin	159999	2.2e-14	reject H_0
red blood cell	148980	2.1e-15	reject H_0

We observe that the p-value of the factors blood pressure, specific gravity, albumin, sugar, blood urea, seram creatinine, sodium, potassium, hemoglobin and red blood cell is lesser than 0.05. Thus, we do reject the null hypothesis. Hence we conclude that the median blood pressure, specific gravity, albumin, sugar, blood urea, seram creatinine, sodium, potassium, hemoglobin and red blood cell of patients having kidney disease is not equal to the patients not having kidney disease.

3.4.2 Analysis of factors affecting the kidney disease using Two proportion Z-test

Assumptions of Two proportion Z-test are,

1. The two samples must be randomly sampled from the two populations.
2. Two samples are independent of each other .
3. Two proportion should be normally distributed. However, for large sample sizes(over 30) this doesn't always matters.

The hypothesis of two proportion Z-test are as follows:

H_0: The proportions are same
H_1: The proportions are not same

Analysis of factor hypertension affecting the presence of kidney disease

The following table shows presence of kidney disease based on presence of hypertension in patients.

Table 10: Contingency table of hypertension and kidney disease

hypertension	Kidney Disease	
	notckd	ckd
no	150	103
yes	0	147

The obtained values are as follows,

- Pearson's chi-squared test statistic = 136.93
- Degrees of freedom = 1
- p-value = 2.2e-16
- Z-score = 11.70171

From the above result, the obtained z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportions are not same and there is a significant difference. Hence we can conclude that, hypertension of the kidney disease patients and not having kidney disease patients are not same.

Analysis of factor diabetes affecting the presence of kidney disease

The following table shows presence of kidney disease based on presence of diabetes in patients.

Table 11: Contingency table of diabetes and kidney disease

Diabetes	Kidney Disease	
	notckd	ckd
no	150	113
yes	0	137

The obtained values are as follows,

- Pearson's chi-squared test statistic = 122.6
- Degrees of freedom = 1
- p-value = 2.2e-15
- Z-score = 11.07249

From the above result, the obtained z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportions are not same and there is a significant difference. Hence we can conclude that, diabetes of the kidney disease patients and not having kidney disease patients are not equal.

Analysis of factor cad affecting the presence of kidney disease

The following table shows presence of kidney disease based on presence of coronary artery disease in patients.

Table 12: Contingency table of coronary artery disease and kidney disease

cad	Kidney Disease	
	notckd	ckd
no	150	216
yes	0	34

The obtained values are as follows,

- Pearson's chi-squared test statistic = 20.581
- Degrees of freedom = 1
- p-value = 2.858e-06
- Z-score = 4.536629

From the above result, the obtained z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportions are not same and there is a significant difference. Hence we can conclude that, coronary artery disease present in patients having kidney disease and not having kidney disease are not equal.

Analysis of factor appetite affecting the presence of kidney disease

The following table shows presence of kidney disease based on presence of appetite in patients.

Table 13: Contingency table of appetite symptom and kidney disease

Appetite	Kidney Disease	
	notckd	ckd
poor	1	82
good	149	168

The obtained values are as follows,

- Pearson's chi-squared test statistic = 56.928
- Degrees of freedom = 1
- p-value = 1.213
- Z-score = 7.545065

From the above result, the obtained z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportions are not same. Hence we can conclude that, appetite present in patients having kidney disease and not having kidney disease are not equal.

Analysis of factor pedel edema affecting the presence of kidney disease

The following table shows presence of kidney disease based on presence of pedel edema in patients.

Table 14: Contingency table of pedel edema symptom and kidney disease

pedel edema	Kidney Disease	
	notckd	ckd
no	150	174
yes	0	76

The obtained values are as follows,

- Pearson's chi-squared test statistic = 54.338
- Degrees of freedom = 1
- p-value = 8.439e-14
- Z-score = 7.371431

From the above result, the obtained z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportions are not same and there is a significant difference. Hence we can conclude that, pedel edema present in patients having kidney disease and not having kidney disease are not equal.

Analysis of factor anemia affecting the presence of kidney disease

The following table shows presence of kidney disease based on presence of anemia in patients.

Table 15: Contingency table of anemia and kidney disease

Anemia	Kidney Disease	
	notckd	ckd
no	150	190
yes	0	60

The obtained values are as follows,

- Pearson's chi-squared test statistic = 40.492
- Degrees of freedom = 1
- p-value = 9.874e-11
- Z-score = 6.363332

From the above result, the obtained z-score is greater than z-table value(1.96). Thus, the null hypothesis is rejected. Therefore, two proportions are not same and there is a significant difference. Hence we can conclude that, anemia present in patients having kidney disease and not having kidney disease are not equal.

Conclusion:

1. From the study of factors affecting the kidney disease using Mann-whitney U test, we observe that median blood pressure, specific gravity, albumin, sugar, blood urea, seram creatinine, sodium, potassium, hemoglobin and red blood cell of patients having kidney disease is not equal to the patients not having kidney disease.

2. By using two proportion Z-test to study about the factors, we observe that the proportion of hypertension, diabetes, coronary artery disease, pedel edema, appetite and anemia of the patients having kidney disease are not equal to the patients not having kidney disease.

3.5 Multivariate analysis of factors affecting the presence of kidney disease

3.5.1 To determine the factors affecting the presence of kidney disease using Logistic Regression.

(a) **Testing the association between Coronary artery disease of the patients and presence of the Kidney Disease using Fisher's exact test of Independence**

The following table shows the classification of the kidney disease patients and coronary artery disease of the respondents.

Table 16: Contingency table on class and coronary artery disease

Class cad	ckd	notckd
no	214	150
yes	34	0

The Hypothesis are as follows:

H_0 : The variables Coronary artery disease and classification are independent.
H_1 : The variables Coronary artery disease and classification are dependent.

The obtained values are as follows:

- p-value = 7.483e-08
- Cramer V = 0.2361

Here we can observe that the obtained p-value is less than 0.05. Hence we do reject the null hypothesis. And we conclude that the variables Coronary artery disease and classification are dependent.

Therefore, there is some relationship between Coronary artery disease of the patients and classification of the kidney disease patients. From the Cramer V value, we conclude that the strength of the association between coronary artery disease and classification is 0.2361. Hence, we conclude that there is a weak association between Coronary artery disease and classification.

(b) **Testing the association between Anemia disease of the patients and presence of the Kidney Disease using Fisher's exact test of Independence**

The following table shows the classification of the kidney disease patients and anemia of the respondents.

Table 17: Contingency table on class and anemia disease

Class ane	ckd	notckd
no	190	150
yes	60	0

The Hypothesis are as follows:

H_0 : The variables Anemia and classification are independent.
H_1 : The variables Anemia and classification are dependent.

The obtained values are as follows:

- p-value = 2.978e-14
- Cramer V = 0.3254

Here we can observe that the obtained p-value is less than 0.05. Hence we do reject the null hypothesis. And we conclude that the variables Anemia and classification are dependent.

Therefore, there is some relationship between Anemia of the patients sand classification of the kidney disease patients. From the Cramer V value, we conclude that the strength of the association between anemia and classification is 0.3254. Hence, we conclude that there is a weak association between anemia and classification.

(c) **Testing the association between pedel edema symptom of the patients and presence of the Kidney Disease using Fisher's exact test of Independence**

The following table shows the classification of the kidney disease patients and pedel edema of the respondents.

Table 18: Contingency table on class and pedel edema disease

Class pe	ckd	notckd
no	174	150
yes	76	0

The Hypothesis are as follows:

H_0 : The variables pedel edema and classification are independent.
H_1 : The variables pedel edema and classification are dependent.

The obtained values are as follows:

- p-value = 2.2e-16
- Cramer V = 0.3752

Here we can observe that the obtained p-value is less than 0.05. Hence we do reject the null hypothesis. And we conclude that the variables pedel edema and classification are dependent.

Therefore, there is some relationship between pedel edema of the patients and classification of the kidney disease patients. From the Cramer V value, we conclude that the strength of the association between pedel edema symptom and classification is 0.3752. Hence, we conclude that there is a weak association between pedel edema and classification.

(d) **Testing the association between diabetes mellitus of the patients and presence of the Kidney Disease using Fisher's exact test of Independence**

The following table shows the classification of the kidney disease patients and diabetes of the respondents.

Table 19: Contingency table on class and diabetes of patients

Class dm	ckd	notckd
no	113	150
yes	137	0

The Hypothesis are as follows:

H_0 : The variables diabetes mellitus and classification are independent.

H_1 : The variables diabetes mellitus and classification are dependent.

The obtained values are as follows:

- p-value = 0.0002
- Cramer V = 0.5591

Here we can observe that the obtained p-value is less than 0.05. Hence we do reject the null hypothesis. And we conclude that the variables diabetes mellitus and classification are dependent.

Therefore, there is some relationship between diabetes mellitus of the patients and classification of the kidney disease patients. From the Cramer V value, we conclude that the strength of the association between diabetes and classification is 0.5591. Hence, we conclude that there is a strong association between diabetes and classification.

Logistic regression model

Let the class of kidney disease patients be considered as the response variable. Here, the predictor variables age, pedel edema, pot, sc, rc, cad, su, pcc, ane, sod, appet, bp, sg, al, ba, bu, hemo, htn and dm of the respondents are taken into consideration to build the model.

The frequency distribution of the response variable as follows,

Table 20: Frequency table for class of kidney disease

class of kidney disease	Frequency
ckd	250
notckd	150

We observe that the data is imbalanced. So, before building the logit model, we must balance the data. let us use up sampling method to balance the data.

The following results are obtained:

Table 21: Summary of logistic regression model

Predictors	Estimate	Std Error	z value	p-value
age	0.03048	0.01702	1.791	0.073311
pedel edema	18.47049	1646.77854	0.011	0.991051
pot	0.73105	0.51130	1.430	0.152774
sc	0.75012	0.26107	2.873	0.004063
rc	-3.98001	0.87930	-4.526	0.000006
cad	12.54789	2517.35	0.005	0.996023
su	6.54051	1.78731	3.659	0.00025
pcc	3.69938	2.70668	1.367	0.1717
ane	18.2623	1828.754	0.010	0.9920
sod	-0.23496	0.05973	-3.934	0.00008

From the above table, we can see the predictor variables, their standard errors, the z-statistic and the associated p-values.

Here the predictor variables ‘sc’,‘rc’,‘su’ and ‘sod’ are statistically significant, since its corresponding p-values are less than 0.05. Hence we can conclude that predictors ‘sc’,‘rc’,‘su’ and ‘sod’ are impact on the kidney disease in patients.

The remaining obtained values are as follows:

Table 22: Results of logistic regression model

Predictors	Estimate	Std Error	z value	p-value
appet	-3.294	1.3940	-2.363	0.0181
bp	0.01768	0.05600	0.316	0.75220
sg	-550.79251	211.85248	-2.600	0.00933
al	2.08718	0.82857	2.519	0.01177
ba	-0.52643	13.62765	-0.039	0.96919
bu	-0.01849	0.04839	-0.382	0.70239
hemo	-2.51451	0.78054	-3.222	0.00128
htn	18.99883	4466.13807	0.004	0.99661
dm	20.60365	4716.87294	0.004	0.99651

From the table we can observe that the p-value of ‘appet’,‘sg’, ‘al’ and ‘hemo’ are less than 0.05. Therefore, these predictor variables are statistically significant. Hence we can conclude that, the predictor variables ‘appet’, ‘sg’, ‘al’ and ‘hemo’ are affecting the kidney disease in patients.

The performance of the model is as follows:

- Train accuracy = 97.9412%
- Test accuracy = 96.6667%
- Root mean square error = 0.1825
- F1-Score = 0.973

We observe that the model performance is good on both train and test sets. F1-Score is also ranges between 0 to 1, and it is also good.

3.6 Decision Tree Classifier

We use decision tree for classifying whether the patient is having kidney disease or not. Classification is considered to be the response variable. Age, bp, sg, al, su, pcc, ba, bu, sc, sod, pot, hemo, rc, htn, dm, cad, appet, pe and ane are used as covariates.

The following decision tree is formed based on CART algorithm where each node is based on the entropy. The R package rpart does this work of building the decision tree and the model with least complexity parameter CP is chosen as the decision tree model.

Here the below table gives the complexity parameter values for different splits.

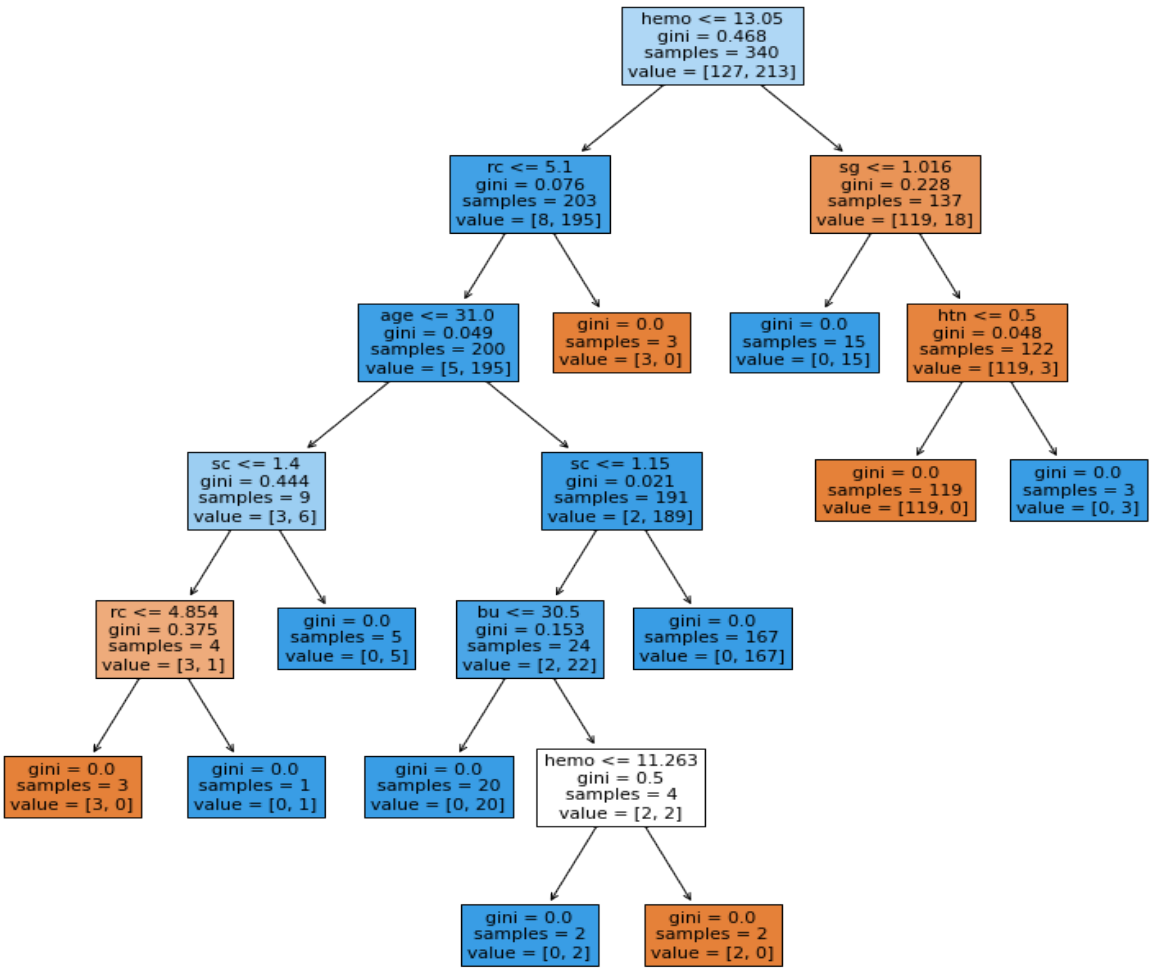
CP	nsplit	Resubstitution error
0.79279	0	1.000000
0.12613	1	0.207207
0.01000	2	0.081081

The performance of the model is as follows:

- Train accuracy = 94.2802%
- Test accuracy = 96.3333%
- Root mean square error = 0.1928
- F1-Score = 0.968

From the above obtained values, We observe that the model performance is good on both train and test sets. F1-Score is also higher. Hence we can conclude this model is a good fit.

Figure 2: Plot of decision tree



From the tree we observe that, gini impurity is high for hemoglobin variable. Thus the first node checks whether hemoglobin is lessar than or equal to 13.05. We got the result that, 340 patients have hemoglobin less than 13.05. In that 127 patients have kidney disease and 213 patients not have kidney disease. In the sample of 340 patients, again it splits into two nodes based on gini impurity. In the left child node, it checks whether the red blood cell is less than 5.1. And we observe that 203 patients have red blood cell less than 5.1, in that 8 patients have kidney disease

and 195 patients not have kidney disease. In the right child node of first node, it checks whether specific gravity less than 1.016. Thus the patients having less specific gravity are 137. In that 119 patients have kidney disease and 18 patients not have kidney disease. Similarly it checks the condition for all the covariates.

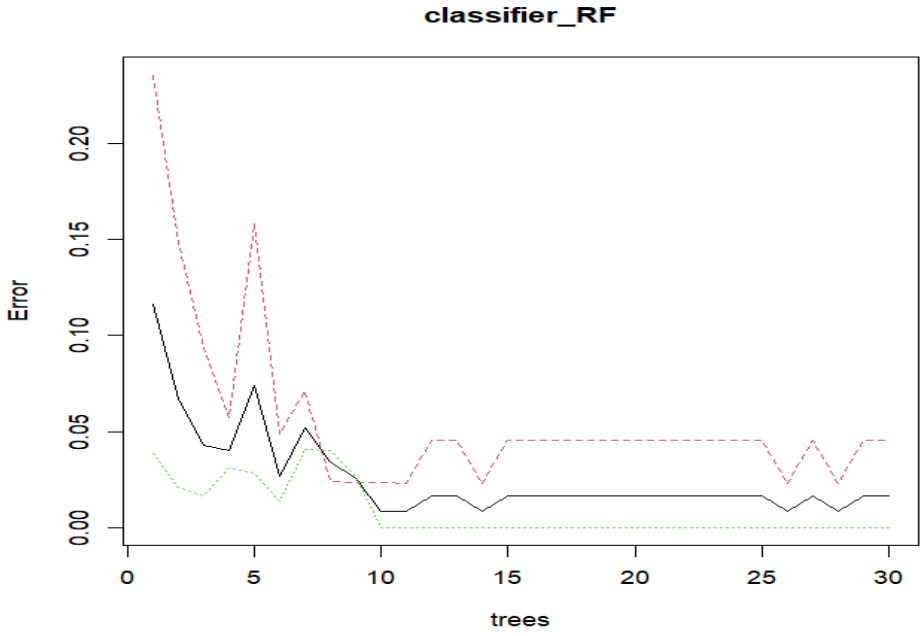
3.7 Random Forest Classifier

Let the class of kidney disease patients be considered as the response variable. Here, the regressor variables age, pedel edema, pot, sc, rc, cad, su, pcc, ane, sod, appet, bp, sg, al, ba, bu, hemo, htn and dm of the respondents are taken into consideration to build the model.

Here the Random Forest classifier is used to build the model to predict whether the patient having the kidney disease or not.

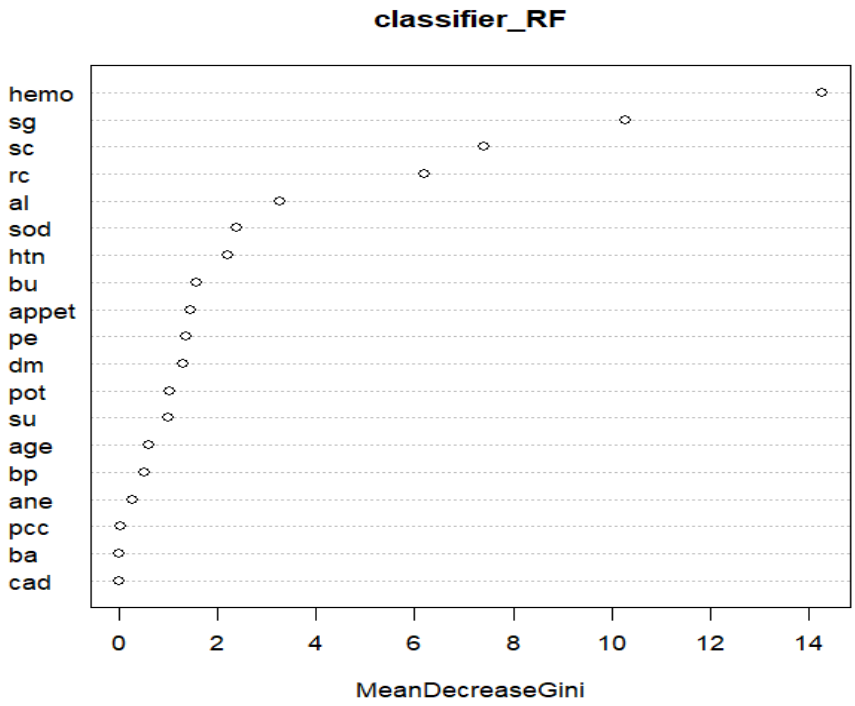
The below plots shows error versus trees in random forest.

Figure 3: Plot of error versus trees in random forest



The below plot shows importance plot of random forest classifier.

Figure 4: Importance plot based on MeandecreaseGini



The mean decrease in Gini coefficient is a measure of how much each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The points in the above plot represents the mean decrease gini value for each variable. The higher the value of mean decrease Gini score, the higher importance of the variable in the model. From the plot we observe that, higher importance is given to hemoglobin variable based on gini coefficient.

The performance of the model is as follows:

- Train accuracy = 98.571%
- Test accuracy = 97.5%
- Root mean square error = 0.0988
- F1-Score = 0.9803

From the above model performance, we conclude that train and test performance is good compare to other models. F1-Score is also very closer to 1, so the model is better. There is no underfit and overfit present in the data.

Model comparision

The below table shows the comparision of all the models.

Table 23: Comparision of all the models

Models	**Accuracy**	**Rmse**	**F1-Score**
Logistic Regression	96.66%	0.1825	0.973
Decision Tree	96.33%	0.1928	0.968
Random Forest	97.5%	0.0988	0.9803

By comparing the performance of all these models, we observe that random forest gives better performance compare to all other models. Hence, we conclude that random forest is the best model to predict the presence of kidney disease in patients.

4 Chapter 4

4.1 Conclusion and Summary

4.1.1 Conclusion

The following are the overall conclusions:

1. The patients age group between 25-64 are having kidney disease in highest number compare to other groups. Hence, we conclude that adults are having highest chance of getting kidney disease.
2. The blood pressure, specific gravity, albumin, sugar, blood urea, seram creatinine, sodium, potassium, hemoglobin and red blood cell of patients having kidney disease is not same for patients not having kidney disease.
   The symptoms of kidney disease like hypertension, diabetes, coronary artery disease, pedel edema and anemia are seem to be higher in patients having kidney disease. Also appetite is good in patients having kidney disease and patients not having kidney disease. Hence the proportion of hypertension, diabetes, coronary artery disease, pedel edema, appetite and anemia of the patients having kidney disease are not equal to patients who are not having kidney disease.
3. The symptom appetite and presence of kidney disease are dependent. Therefore, we conclude that there is a relationship between appetite symptom and presence of kidney disease.
   Also we observe that hypertension disease and presence of kidney disease are dependent. Therefore we conclude that, there is a association between hypertension of the patients and presence of kidney disease.
4. The factors serum creatinine which is left in muscles, red blood cells, sugar level, sodium content, specific gravity of the urine, albumin, hemoglobin and the appetite symptom is impacting on the kidney disease in patients.
5. The performance of random forest model is good. So, we use random forest model to predict the presence of kidney disease.

4.1.2 Summary

A project entitled "Predictive Analysis on Chronic Kidney Disease" has been done. A secondary data has been collected from the website of UCI machine repository. The subjects considered in this study are patients aged from 0 to 100 years with problem of kidney disease. This data is collected in the year 2015. The dataset consists of 400 observations and 21 variables.

The analysis and interpretation of the data is done by using some of the statistical methods like Logistic regression, Fisher's exact test of independence, Mann-Whitney U test, Two proportion z-test, Decision tree classifier, Random forest classifier and some of the visualization techniques.

From the results obtained, we came to know that adults are having highest chance of getting kidney disease compare to childrens, youths and elders. There is a weak relationship between the appetite symptom and presence of kidney disease. And there is a strong association between hypertension of the patients and presence of kidney disease. The median blood pressure, specific gravity, albumin, sugar, blood urea, seram creatinine, sodium, potassium, hemoglobin and red blood cell of patients having kidney disease is not same for patients not having kidney disease. Also the proportion of hypertension, diabetes, coronary artery disease, pedel edema, appetite and anemia of the patients having kidney disease are not equal to patients who are not having kidney disease. The factors like serum creatinine, red blood cells, sugar level, sodium content, specific gravity, albumin, hemoglobin and the appetite symptom is affecting on the kidney disease in patients. Finally by comparing the performance of Logistic regression, Decision tree and Random forest model, we observe that random forest model performance is best fit to predict the presence of kidney disease in patients.

5 Chapter 5

5.1 Bibliography

1. www.statstest.com/two-proportion-z-test
2. www.statisticshowto.com/fishers-exact-test-independence
3. www.sciencedirect.com/topics/biochemistry-genetics-and- molecular-biology/mann-whitney-u-test
4. www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
5. www.sthda.com/english/articles/36-classification-methods-essentials/ 151-logistic-regression-essentials-in-r
6. www.thestatsgeek.com/2014/04/12/is-the-wilcoxon-mann-whitney-test- a-good-non-parametric-alternative-to-the-t-test
7. www.monkeylearn.com/blog/classification-algorithms
8. www.scikit-learn.org/stable/modules/generated/sklearn.ensemble. RandomForest-Classifier.html
9. Ifraz, G. M. (2021). *Comparative Analysis for Prediction of Kidney Disease using Intelligent Machine Learning methods.* National Health and Nutrition Examination Survey. University of MICHIGAN.
10. Revathy, S. and Ramesh, M. (2019). *Predicting the presence Chronic kidney Disease using models.* Screening for Occult Renal Disease(SCORED).
11. Krishamurthy, S. (2021). *Chronic Kidney Disease using National Health Insurance Claim Data.* Findings from the National Health and Nutrition Survey(NHANES).
12. Gulati, A. P. (2022). *Performance Evaluation of different Machine learning Classification Algorithm for Disease Diagnosis.* American University.
13. Chittora, P. (2021). *Risk factors for progressive chronic kidney disease.* K/DOQI clinical practice guildlines for chronic kidney disease.
14. Mienye, D. I. (2022). *Prediction of Chronic Kidney Disease Using Feature Selection and Boosted Classifiers.* New Engl J Med.
15. Mohan, V. (2015). *Kidney Disease Prediction Using SVM and ANN algorithm.* Review. Am J Kidney Dis.
16. Saini, J. (2016). *Design and Analysis of algorithm for prediction of Kidney Disease.* Soc Nephrol(Volume 4).

6 Chapter 6

6.1 Appendix

R CODES

```
#importing data
data=read.csv("LR data.csv")
head(data)

#fishers exact test of independence
chi=table(data$appet,data$classification)
chi
chisq.test(chi)$expected
fisher.test(chi)
library(rcompanion)
cramerV(chi)

chi=table(data$htn,data$classification)
chi
chisq.test(chi)$expected
fisher.test(chi)
cramerV(chi)

chi=table(data$dm,data$classification)
chi
chisq.test(chi)$expected
fisher.test(chi)
cramerV(chi)

chi=table(data$cad,data$classification)
chi
chisq.test(chi)$expected
fisher.test(chi)
cramerV(chi)

chi=table(data$ane,data$classification)
chi
```

```
chisq.test(chi)$expected
fisher.test(chi)
cramerV(chi)

chi=table(data$pe,data$classification)
chi
chisq.test(chi)$expected
fisher.test(chi)
cramerV(chi)

#QQ-plot
library(ggpubr)
ggqqplot(data,"bp",facet.by="classification")
ggqqplot(data,"sg",facet.by="classification")
ggqqplot(data,"al",facet.by="classification")
ggqqplot(data,"su",facet.by="classification")
ggqqplot(data,"bu",facet.by="classification")
ggqqplot(data,"sc",facet.by="classification")
ggqqplot(data,"sod",facet.by="classification")
ggqqplot(data,"pot",facet.by="classification")
ggqqplot(data,"hemo",facet.by="classification")
ggqqplot(data,"rc",facet.by="classification")

#Anderson Darling normality test
library(nortest)
tapply(data$bp,data$classification,ad.test)
tapply(data$sg,data$classification,ad.test)
tapply(data$al,data$classification,ad.test)
tapply(data$su,data$classification,ad.test)
tapply(data$bu,data$classification,ad.test)
tapply(data$sc,data$classification,ad.test)
tapply(data$sod,data$classification,ad.test)
tapply(data$pot,data$classification,ad.test)
tapply(data$hemo,data$classification,ad.test)
tapply(data$rc,data$classification,ad.test)

#Mann-Whitney U test
wilcox.test(data$bp,data$classification)
```

```
wilcox.test(data$sg,data$classification)
wilcox.test(data$al,data$classification)
wilcox.test(data$su,data$classification)
wilcox.test(data$bu,data$classification)
wilcox.test(data$sc,data$classification)
wilcox.test(data$sod,data$classification)
wilcox.test(data$pot,data$classification)
wilcox.test(data$hemo,data$classification)
wilcox.test(data$rc,data$classification)

#Two proportion z-test
a)
data['classification']=lapply(data['classification'], factor,
                        levels=c(0,1),labels=c("notckd","ckd"))
data['htn']=lapply(data['htn'], factor, levels=c(0,1),labels=c("no","yes"))
dt=table(data$htn,data$classification)
dt

barplot(dt,ylab='No. of patients having hypertension',beside=TRUE)
prop.test(dt,alternative="greater")
z=sqrt(136.93)
z

b)
data['dm']=lapply(data['dm'], factor, levels=c(0,1),
           labels=c("no","yes"))
dt=table(data$dm,data$classification)
dt

barplot(dt,ylab='Count of diabetes patient',beside=TRUE)
prop.test(dt,alternative="greater")
z=sqrt(122.6)
z

c)
data['cad']=lapply(data['cad'], factor, levels=c(0,1),
            labels=c("no","yes"))
dt=table(data$cad,data$classification)
```

```
dt

barplot(dt,ylab='No. of patients having artery disease',beside=TRUE)
prop.test(dt,alternative="greater")
z=sqrt(20.581)
z

d)
data['appet']=lapply(data['appet'], factor, levels=c(0,1),
              labels=c("poor","good"))
dt=table(data$appet,data$classification)
dt

barplot(dt,ylab='Count of patients having Appetite',beside=TRUE)
prop.test(dt,alternative="greater")
z=sqrt(56.928)
z

e)
data['pe']=lapply(data['pe'], factor, levels=c(0,1),labels=c("no","yes"))
dt=table(data$pe,data$classification)
dt

barplot(dt,ylab='Count of patients having pedel edema',beside=TRUE)
prop.test(dt,alternative="greater")
z=sqrt(54.338)
z

f)
data['ane']=lapply(data['ane'], factor, levels=c(0,1),labels=
            c("no","yes"))
dt=table(data$ane,data$classification)
dt

barplot(dt,ylab='Count of patients having anemia',beside=TRUE)
prop.test(dt,alternative="greater")
z=sqrt(40.492)
z
```

```
#z-table value
ztab=qnorm((1-(0.05/2)),0,1)
ztab


# Random forest model
data$classification=factor(data$classification,levels=c(0,1))
data$htn=factor(data$htn,levels=c(0,1))
data$dm=factor(data$dm,levels=c(0,1))
data$cad=factor(data$cad,levels=c(0,1))
data$appet=factor(data$appet,levels=c(0,1))
data$ane=factor(data$ane,levels=c(0,1))
data$pe=factor(data$pe,levels=c(0,1))
data$pcc=factor(data$pcc,levels=c(0,1))
data$ba=factor(data$ba,levels=c(0,1))

# Check number of rows and columns
dim(data)

str(data)

library(caTools)
library(randomForest)

#splitting data in train and test data
split=sample.split(data,SplitRatio=0.7)

train=subset(data,split=="TRUE")
test=subset(data, split=="FALSE")
head(test)
head(train)

dim(train)
dim(test)

#Fitting random forest to the train data
set.seed(120)
```

```
classifier_RF=randomForest(x=test[-20],y=test$classification,
              importance=TRUE)
summary(classifier_RF)
varImpPlot(classifier_RF)

#Predicting the test set results
y_pred_test=predict(classifier_RF,newdata=test[-20])
y_pred_train=predict(classifier_RF,newdata=train[-20])

#confusion matrix
Confusion_Matrix=table(test[,20],y_pred_test)
Confusion_Matrix

#plotting model
plot(classifier_RF)

#importance plot
importance(classifier_RF)

accuracy=(103+173)/(103+173+3+1)#train
accuracy

accuracy=(42+75)/(75+42+1+2)#test
accuracy

precision=75/(75+1)
recall=75/(75+2)
F1_score=2*recall*precision/(recall+precision)
F1_score


# Logistic regression model
table(data$classification)
library(caret)
'%ni%'=Negate('%in%')
options(scipen=999)
```

```
set.seed(100)
trainDataIndex=createDataPartition(data$classification,p=0.7,list=FALSE)

trainData=data[trainDataIndex,]

testData=data[-trainDataIndex,]

table(trainData$classification)

#up sample
set.seed(100)
up_train <- upSample(x=trainData[, colnames(trainData) %ni%
             "classification"],
              y=trainData$classification)
up_train
table(up_train$Class)

# Build Logistic Model
logitmod <- glm(Class ~ ., family = "binomial", data=up_train)
summary(logitmod)

pred <- predict(logitmod, newdata = testData, type = "response")
pred
y_pred_num <- ifelse(pred > 0.5, 1, 0)
y_pred <- factor(y_pred_num, levels=c(0, 1))
y_act <- testData$classification
mean(y_pred == y_act) #accuracy




# Decision tree model
# Split data into training (70%) and validation (30%)
dt = sort(sample(nrow(data), nrow(data)*.7))
train<-data[dt,]
train
dim(train)
val<-data[-dt,] # Check number of rows in training data set
```

```
val
dim(val)
nrow(train)

#to view the dataset
edit(train)

# Decision Tree Model
library(rpart)
mtree <- rpart(classification~., data = train, method="class",
        control = rpart.control(minsplit = 20, minbucket = 7,
maxdepth = 10, usesurrogate = 2, xval =10 ))
mtree

#plot the tree
plot(mtree)
text(mtree)

#Beautify tree
library(rattle)
library(rpart.plot)
library(RColorBrewer)

#view1
prp(mtree, faclen = 0, cex = 0.8, extra = 1)

#view2 - total count at each node
tot_count <- function(x, labs, digits, varlen)
{paste(labs, "\n\nn =", x$frame$n)}

prp(mtree, faclen = 0, cex = 0.8, node.fun=tot_count)

#view3- fancy Plot
library(rattle)
fancyRpartPlot(mtree)

printcp(mtree)
bestcp <- mtree$cptable[which.min(mtree$cptable[,"xerror"]),"CP"]
```

```
# Prune the tree using the best cp.
pruned <- prune(mtree, cp = bestcp)

# Plot pruned tree
prp(pruned, faclen = 0, cex = 0.8, extra = 1)

# confusion matrix (training data)
conf.matrix=table(train$classification, predict(pruned,type="class"))
rownames(conf.matrix)=paste("Actual", rownames(conf.matrix), sep = ":")
colnames(conf.matrix)=paste("Pred", colnames(conf.matrix), sep = ":")
print(conf.matrix)

#Scoring
library(ROCR)
val1 = predict(pruned, val, type = "prob")
#Storing Model Performance Scores
pred_val <-prediction(val1[,2],val$classification)

# Calculating Area under Curve
perf_val <- performance(pred_val,"auc")
perf_val

# Plotting Lift curve
plot(performance(pred_val, measure="lift", x.measure="rpp"), colorize=TRUE)

# Calculating True Positive and False Positive Rate
perf_val <- performance(pred_val, "tpr", "fpr")

# Plot the ROC curve
plot(perf_val, col = "green", lwd = 1.5)

#Calculating KS statistics
ks1.tree <- max(attr(perf_val, "y.values")[[1]] - (attr(perf_val,
            "x.values")[[1]]))
ks1.tree
```

```
# Advanced Plot
prp(pruned, main="Beautiful Tree",
    extra=106,
    nn=TRUE,
    fallen.leaves=TRUE,
    branch=.5,
    faclen=0,
    trace=1,
    shadow.col="gray",
    branch.lty=3,
    split.cex=1.2,
    split.prefix="is ",
    split.suffix="?",
    split.box.col="lightgray",
    split.border.col="darkgray",
    split.round=.5)
```