Exploring Kidney Disease through Computer Simulations: Analyzing Missing Data, Associations, and Classification Techniques

-Shrinidhi Rajesh

Abstract:
In the intricate landscape of kidney disease, this study delves deep into a comprehensive dataset, unveiling hidden characteristics, revealing intricate associations, and shedding light on effective classification, igniting the path towards enhanced diagnosis and treatment. Kidney disease is a prevalent and significant health concern, and understanding its factors and patterns is crucial for effective diagnosis and treatment. This project aims to enhance the analysis and interpretation of a kidney disease dataset by addressing missing data, exploring associations, and improving classification techniques. Missing data pose a common challenge in healthcare research, potentially impacting patient care and research outcomes. By employing appropriate imputation techniques, the study focuses on creating a complete dataset to enhance accuracy and reliability. Additionally, bootstrap resampling techniques was used to estimate population parameters and to compare variable means between patients with and without chronic kidney disease (CKD), shedding light on key differences. The application of permutation testing provides a robust assessment of correlations, uncovering statistically significant associations and distinguishing them from random chance. Furthermore, the study utilizes linear discriminant analysis and Monte Carlo simulations for classification purposes, evaluating the performance of a model in predicting CKD. Feature selection techniques, including permutation testing, contribute to identifying important predictors for kidney disease classification. Finally, logistic regression models with all the attributes and the selected attributes through feature selections were compared to assess their predictive accuracy. This comprehensive analysis contributes to advancing knowledge in the field of chronic kidney disease research and improving diagnostic and treatment strategies.

Table Of Contents:

Introduction:

Chronic kidney disease (CKD) is a significant health concern worldwide, characterized by the gradual loss of kidney function over time. It is associated with various risk factors, including age, hypertension, diabetes mellitus, and other comorbidities. Early detection and accurate diagnosis of CKD are crucial for effective management and treatment to prevent its progression and associated complications. In this study, a comprehensive dataset comprising blood tests and other measures from patients with and without CKD was utilized. The dataset consists of 400 rows, representing individual patients seen over a period of approximately two months before July 2015 in a hospital located in Tamil Nadu, India, possibly Apollo Reach Karaikudi. Out of the 400 rows, 250 correspond to patients diagnosed with CKD, while the remaining 150 rows correspond to patients without CKD. The classification information is provided in the "classifcation" column of the dataset. To gain a comprehensive understanding of CKD, it is essential to explore the various factors and patterns associated with the disease. The dataset includes a range of variables that provide valuable insights into patient characteristics and potential risk factors. These variables encompass information such as age, as well as measurements related to blood pressure, specific gravity of urine, albumin and sugar levels in urine, presence of abnormal red blood cells and pus cells in urine, presence of pus cell clumps and bacteria in urine, blood glucose levels, blood urea and serum creatinine concentrations, sodium and potassium levels, hemoglobin and packed cell volume, white and red blood cell counts, and the presence of hypertension, diabetes mellitus, coronary artery disease, poor appetite, pedal edema, and anemia.

Existing literature emphasizes the importance of laboratory measurements and diagnostic parameters in assessing kidney function and diagnosing CKD. Studies by Vassalotti et al. (2007) highlight the significance of blood tests and urine analysis, including parameters such as blood pressure, urine specific gravity, albumin, sugar levels, and presence of abnormal cells and bacteria, in evaluating kidney health and identifying CKD. According to the Indian Chronic Kidney Disease Registry (ICKD), the prevalence of CKD in India has been steadily increasing over the years. The ICKD is a nationwide initiative that aims to collect and analyze data on CKD prevalence, risk factors, and treatment outcomes. Their findings indicate that CKD has become a significant public health issue in India, with a high burden of cases observed across various regions of the country. A study conducted by Jha et al. (2013) in India estimated the prevalence of CKD to be around 17.2% in the general population. The study emphasized the growing burden of CKD in India and highlighted the need for early detection and intervention to prevent disease progression and complications. The impact of comorbid conditions on CKD has also been investigated in the Indian context. A study by Agarwal et al. (2016) explored the association between CKD and comorbidities such as hypertension and diabetes mellitus in a rural population in India. The results showed a high prevalence of hypertension and diabetes among CKD patients, suggesting a strong correlation between these conditions and the development of CKD in the Indian population. Furthermore, a study by Joshi et al. (2015) examined the association between CKD and cardiovascular diseases, including coronary artery disease, in Indian patients. The findings highlighted the higher risk of cardiovascular complications in individuals with CKD and

emphasized the need for integrated management strategies that address both CKD and associated comorbidities.
These studies underscore the importance of early detection, effective management of comorbidities, and comprehensive healthcare approaches in the Indian context to tackle the rising burden of CKD. The inclusion of variables related to hypertension, diabetes mellitus, and coronary artery disease in the dataset being analyzed in this study allows for a deeper exploration of their impact on CKD and provides insights into the specific context of CKD in India. By building upon the existing literature and incorporating the Indian perspective, this study aims to conduct a comprehensive analysis of this dataset by contributing to the existing knowledge in the field of CKD research, improving diagnostic strategies, and enhancing the treatment approaches. The findings of this study can potentially inform healthcare professionals in their decision-making process, leading to better patient care and improved outcomes for individuals affected by CKD, especially in India.

Purpose of the Study:

1) The very first goal of this study was to handle missing data by using appropriate imputation techniques. By addressing missing values and creating a complete dataset, the aim was to enhance the accuracy and reliability of subsequent analyses and interpretations. Missing data are very common when dealing with healthcare research and surveys in general, where incomplete or missing information can have significant implications for patient care, treatment decisions, and medical research outcomes. By effectively handling missing data, this analysis aims to contribute to the improvement of data quality and the advancement of knowledge in the field of healthcare especially for the chronic kidney disease research.
2) Bootstrap resampling technique was used to estimate population parameters for the attributes in the kidney disease dataset. These statistics will provide a concise representation of the central tendency and variability of the data, allowing for better understanding and interpretation of the dataset. Furthermore, the application of bootstrap resampling in estimating population parameters contributes to a comprehensive understanding of the kidney disease dataset, enabling researchers and healthcare professionals to make informed decisions based on reliable statistical information.
3) Bootstrap was again utilized to construct Confidence Intervals. The aim of this analysis is to compare the mean values between patients with and without chronic kidney disease is to identify potential differences in various attributes. Understanding these differences can help in detecting patterns, risk factors, or indicators associated with kidney disease. This knowledge can contribute to better diagnosis, treatment, and management of kidney disease in real-life healthcare settings.
4) Permutation testing, also known as randomization testing, is a statistical technique used to assess the statistical significance of an observed effect or relationship in a dataset. It is particularly useful when the underlying distribution of the data is unknown or violates certain assumptions required by traditional parametric tests. Here, the permutation test was used to assess whether the observed correlations in the original data matrix are statistically significant or can be explained by random chance. On the other hand, the

traditional correlation matrix provides the correlations between variables in the original dataset but does not assess their statistical significance. It does not account for the possibility that the observed correlations could be the result of random variation rather than true associations between the variables. By permuting the data and computing correlation matrices based on permuted datasets, the code generates a null distribution of correlations. Comparing the observed correlations in the data matrix with this null distribution allows to evaluate the statistical significance of the correlations. By calculating correlation coefficients and p-values, the aim was to identify significant associations between variables and gain insights into their relationships. Understanding these correlations can help in identifying potential risk factors or indicators for kidney disease and provide valuable information for further research and clinical decision-making. Therefore, permutation testing provides a more robust approach to assess the statistical significance of correlations and helps in distinguishing between meaningful relationships and random chance.

5) Classification of the kidney disease dataset based on the two groups of individuals affected with CKD and the non – CKD individuals was performed. The classification was based on linear discriminant analysis. To evaluate the performance of a Linear Discriminant Analysis (LDA) model on the classification of kidney disease using various predictor variables, Monte Carlo Simulations was performed. Monte Carlo simulation is used here to evaluate the performance of the LDA model due to its ability to generate multiple random samples and assess the model's performance across these samples. By randomly sampling the dataset multiple times and fitting the LDA model on each sampled dataset, a more robust estimation of the model's performance metrics is obtained. Secondly, Monte Carlo simulations were used to analyze the class probabilities predicted by the LDA model for the kidney disease dataset. The usage of Monte Carlo Simulation to evaluate the classification models can be valuable in clinical practice for early detection, risk assessment, and personalized treatment strategies for patients with kidney disease.
6) Feature selection was performed using permutation, to determine the importance of individual features in a logistic regression model for classifying kidney disease. By permuting the values of each feature and maeasuring the decrease in the area under the receiver operating characteristic (AUC-ROC) curve, the impact of each feature on the model's performance is assessed. This knowledge can aid in understanding the underlying mechanisms, identifying biomarkers, and improving diagnostic and prognostic models for kidney disease.
7) Lastly, permutation was again performed to compare the predictive accuracy of two logistic regression models: the full model with only the 8 attributes resulted from the feature selection using permutation testing on the original model. and the original model with all the attributes in the dataset with categorical variables converted as numeric and the attribute "pcv" was removed due to high collinearity with "hemo". By assessing different models, researchers can identify the most effective approach for predicting kidney disease based on the available attributes. This knowledge can assist in the development of more accurate diagnostic models, leading to improved patient outcomes, treatment decisions, and resource allocation in real-life healthcare settings.

8) In summary, each goal serves a specific real-life purpose, ranging from improving data quality, identifying associations and risk factors, developing classification models, selecting influential features, and evaluating predictive accuracy. These purposes ultimately contribute to advancing knowledge in the field of healthcare, improving patient care, and aiding in clinical decision-making related to chronic kidney disease.

Methodology:

The kidney disease dataset was obtained from Kaggle and loaded in R. R was used in this project due to its exemplary inbuilt functions which are robust in handling missing data and performing computer simulations.

The dataset includes 400 instances with 25 attributes with blood tests and other measures from patients with and without CKD. Each instance depicts patients seen over a period of about two months at some point before July 2015, in a hospital in Tamil Nadu, India; Apollo Reach Karaikudi. Of the 400 rows, 250 correspond to patients with CKD and the remaining 150 rows correspond to patients without CKD [Table 1]

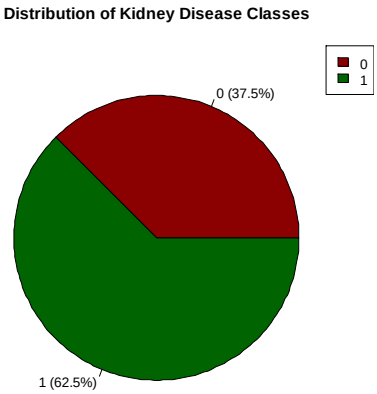

Table 1: Distribution of the Kidney Disease Classes

The summary statistics [Table 2] are calculated and the description of each column along with their description and levels of measurement is tabulated [Table 3].

After loading the dataset, the presence of missing data was assessed by calculating the sum of missing values using the is.na function, which revealed a total of 1012 missing values [Table 4]. To handle the missing data, the R package "mice" was utilized. The "mice" package provides methods for multiple imputation, which is a technique used to estimate missing values by generating value plausible values based on observed data. In this analysis, the "pmm" method (Predictive Mean Matching) was chosen for imputing missing values. The mice function was applied to the dataset, specifying the number of imputations as 1(m=1),

the maximum number of iterations is set to 50 (maxit = 50), the imputation method, and a seed for reproducibility, respectively. Furthermore, a density plot[[Table 5] was created to visualize the distribution of the imputed data. To obtain a single imputed dataset, the first and only set of imputed values were selected using the complete function from the "mice" package. The imputed dataset was then examined again for missing data. It was found that 234 missing values remained in the nominal columns. To handle these missing values, the mode (the most frequent value) of each nominal variable was computed using the table function. The missing values in each nominal column were then replaced with their corresponding mode values using indexing. Finally, the dataset was checked once again for missing values to ensure that all missing data had been appropriately imputed. No missing values were found, indicating that the imputation process was successful [Table 6].

The "mice" package was specifically chosen to handle missing data in this analysis due to its robustness and effectiveness in multiple imputation. Multiple imputation is a technique that generates plausible values for missing data based on observed data, allowing for a more accurate representation of the underlying population. The "pmm" method is particularly useful when dealing with continuous variables as it predicts missing values by finding observed values with similar characteristics. This method helps preserve the underlying distribution and relationships in the data, resulting in more reliable imputations.

To handle the missing values in the nominal/categorical columns, the mode (the most frequent value) of each nominal variable was computed. The mode is a suitable choice for nominal variables because it represents the category that occurs most frequently in the data. By replacing the missing values with the mode values using indexing, the nominal variables were imputed appropriately. This approach ensures that the imputed values align with the most prevalent categories in the dataset and maintains the categorical nature of the variables. The combination of the "mice" package for multiple imputation and the computation of mode values for nominal variables allowed for a comprehensive handling of missing data, ensuring that the dataset was adequately imputed and ready for further analysis.

For easiness of coding, a new column named "classify" was created in the "kidneydisease" dataset. If the value in the "classification" column was "ckd," it was classified as 1; otherwise, it was classified as 0. The resulting binary classification was stored in the "classify" column, which was then printed. To further process the data, several nominal variables were converted to binary format. The "rbc," "pc," "pcc," "ba," "htn," "dm," "appet," "pe," "ane," and "cad" columns were transformed into binary variables as follows:

The "RBC" column was created based on the "rbc" column, where "normal" values were assigned 0 and other values were assigned 1.
The "PC" column was created based on the "pc" column, where "normal" values were assigned 0 and other values were assigned 1.
The "PCC" column was created based on the "pcc" column, where "notpresent" values were assigned 0 and other values were assigned 1.
The "BA" column was created based on the "ba" column, where "notpresent" values were assigned 0 and other values were assigned 1.

The "HTN" column was created based on the "htn" column, where "no" values were assigned 0 and other values were assigned 1.
The "DM" column was created based on the "dm" column, where "no" values were assigned 0 and other values were assigned 1.
The "APPET" column was created based on the "appet" column, where "good" values were assigned 1 and other values were assigned 0.
The "PE" column was created based on the "pe" column, where "no" values were assigned 0 and other values were assigned 1.
The "ANE" column was created based on the "ane" column, where "no" values were assigned 0 and other values were assigned 1.
The "CAD" column was created based on the "cad" column, where "yes" values were assigned 1 and other values were assigned 0.
Additionally, a binary variable named "AGE" was created based on the "age" column. Values less than or equal to 52 were assigned 0, while values greater than 52 were assigned 1.

For all the further analysis, only the attributes "age," "bp," "sg," "al," "su," "bgr," "bu," "sc," "sod," "pot," "hemo," "wc," "rc," "RBC," "PC," "PCC," "BA," "HTN," "DM," "APPET," "PE," "ANE," "CAD," were utilized. All the nominal variables were used in numeric form. The attribute "pcv" was not used in most of the analysis due to its high collinearity with "hemo"[Table 7].
To estimate population parameters for the attributes in the kidney disease dataset the bootstrap resampling was utilized. Bootstrap resampling is a statistical technique that involves repeatedly sampling from the original dataset to create multiple resamples. For each resample, the desired summary statistic [Table 8] in this case, the min, lower_ci median, mean, upper_ci and max are calculated. By averaging these summary statistics across the resamples, an estimate of the population is obtained. Using bootstrap resampling for summary statistics offers several advantages over traditional approaches. Firstly, bootstrap resampling does not rely on any assumptions about the underlying data distribution, making it more robust and applicable to a wide range of datasets. Secondly, by generating multiple resamples, it captures the inherent variability and provides a more accurate estimate of the population characteristics. Lastly, the computation of bootstrap confidence intervals allows for quantifying the uncertainty associated with the estimated summary statistics, offering a more comprehensive understanding of the data.

Proceeding further I performed t-tests to comprehensively assess the significance of the mean difference between the CKD and non-CKD groups.

Hypothesis:

H0: No true difference in the means between the two groups.

Ha: There is a significant difference in the means between the two groups.

It should be noted that the t test was applied exclusively to the discrete and continuous variables, and not to the categorical variables. The t-tests indicated significant differences (p values<0.05) in the means of all variables except for 'pot' between the two groups[Table 9].

The t-test implies that the variable "pot" may not be a strong differentiating factor between CKD and non-CKD cases.

To investigate further and strengthen the findings and ensure a more rigorous analysis, I utilized bootstrap analysis. The goal of this analysis was to determine if there is a significant difference in means between Chronic Kidney Disease (CKD) and non-CKD groups for various variables. The bootstrap technique involves randomly sampling the data with replacement to create multiple bootstrap samples, from which statistics of interest are computed. This bootstrap resampling technique helps to assess the uncertainty and variability in the dataset by calculating bootstrap confidence intervals for the continuous variables in the kidney disease dataset. By applying the bootstrap method, the aim was to obtain more reliable estimates of central tendency and confidence intervals, which can provide valuable insights into the characteristics and distribution of the variables. This goal reflects the importance of understanding the variability in real-life data and its implications for decision-making and statistical inference. By using the bootstrap, the mean difference between the CKD and non-CKD groups for each variable, along with their corresponding confidence intervals were calculated. The methodology involved using the bootstrap method and the mean function to estimate the means of different variables, including age, blood pressure (bp), specific gravity (sg), albumin (al), sugar (su), blood glucose random (bgr), blood urea (bu), and serum creatinine (sc). For each variable, separate bootstrap analyses were performed for the CKD and non-CKD groups, with 1000 resamples. The bootstrap results provided confidence intervals (CI) for each variable in both the CKD and non-CKD groups. To achieve the goal, the R package "boot" was utilized. The continuous attributes in the dataset were stored in a separate entity and only those 14 attributes were subjected to bootstrap resampling seperately. The resampling process involved randomly selecting observations with replacement from the original dataset to create multiple bootstrap samples. For each sample, the mean of the variable was calculated as the statistic of interest. This procedure was repeated a specified number of times (num_samples = 1000) to generate a robust distribution of sample means. Bootstrap confidence intervals were then computed using the boot.ci function, employing the percentile method ("perc"). Here I used boot.ci as my version of R didn’t support the quantile function. The lower and upper bounds of the confidence intervals provided an estimate of the range within which the population parameter (mean) of each variable was likely to fall [Table 10].

To evaluate the correlation between the attributes, a correlation matrix [Table 11] was calculated using the Pearson correlation coefficient for the selected variables. The selected variables, such as age, blood pressure, specific gravity, albumin, sugar, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, and red blood cell count, were used in the analysis due their continuous nature. The correlation matrix was computed using the correlation function. Permutation was performed by setting the number of permutations to 1000, and for each permutation, the data was randomly shuffled, and a new correlation matrix was computed. The final permuted correlation matrix was obtained by averaging the permuted correlation matrices[Table 12]. The p-values were determined by comparing the observed correlations with permuted correlations obtained from randomly shuffled data. To calculate the p-values,

each observed correlation coefficient was compared against the absolute values of permuted correlations, and the proportion of permuted correlations that were greater than or equal to the observed correlation was divided by the total number of permutations. The resulting p-values were then used to assess the significance of the correlations between pairs of variables.[Table 13]

To evaluate the LDA model, MC simulations were performed. The number of Monte Carlo simulations is set to 10,000 (num_simulations = 10000). Empty vectors are created to store the evaluation metrics: accuracy, precision, sensitivity (recall), and F1 score. A matrix is initialized to store the class probabilities. For each simulation, Random sampling with replacement is performed on the kidney disease dataset to create a sampled dataset. The LDA model is fitted on the sampled data based on the proportion of each groups in the dataset. Ckd Group: 250/400 and non ckd group: 150/400 thus the prior probability was set to c(0.375, 0.625)[Table 14]. The class labels are predicted for the original dataset using the trained LDA model. Evaluation metrics (accuracy, precision, sensitivity, and F1 score) are calculated using the predicted labels and actual labels. The evaluation metric values are stored in their respective vectors[Table 15]. The class probabilities are predicted for the original dataset using the trained LDA model. The predicted class probabilities are added to the matrix of class probabilities. The mean and standard deviation of the evaluation metrics are calculated using the stored values from the Monte Carlo simulations. The confusion matrix is generated to show the classification results. In addition to evaluating performance metrics, the class probabilities for each instance in the original dataset are predicted using the trained LDA model. The predicted class probabilities from all simulations are summed and divided by the number of simulations to obtain the average probabilities. The average probabilities are rounded to four decimal places. The class probabilities, along with the actual classification labels and a column indicating correct/incorrect classification, are stored in a matrix. The percentage of correctly classified instances is calculated [Table 16].

To perform Feature selection using permutation, a logistic regression model is trained using all features in the kidney disease dataset. The initial AUC-ROC is calculated for this model. Then, feature permutation is performed by randomly shuffling the values of each feature while keeping the target variable unchanged. For each permuted feature, the logistic regression model is applied, and the AUC-ROC is computed. The decrease in AUC-ROC from the initial value indicates the importance of the permuted feature. This process is repeated for all features, resulting in an importance score for each feature. [Table 17]

Two logistic regression models are fitted: the full model with 8 attributes sg, al, sc, hemo, bu, bgr, HTN, and bp, and the original model with all attributes. Permutation testing is performed by randomly permuting the outcome variable and refitting the models on the permuted data with 1000 simulations. The observed difference in AUC-ROC (Area Under the Receiver Operating Characteristic Curve) is calculated between the full model and the original model. The permutation test p-value is computed by comparing the observed difference with the distribution of permuted differences. Statements are printed based on the AUC-ROC comparison and permutation test results.[Table 18]

Discussion and Results:

The results of the bootstrap resampling provide valuable insights into the dataset. For each variable of interest, we obtain summary statistics including the minimum value, lower confidence interval, median, mean, upper confidence interval, and maximum value. These statistics offer a comprehensive understanding of the central tendency and spread of each variable, allowing for comparisons and identifying potential outliers. By incorporating bootstrap analysis, I compared if there is overlap in the confidence interval between the ckd and non ckd groups as well as if 0 exist in the Confidence interval. The result provided additional evidence that supports the significant mean difference between the CKD and non-CKD groups for all the 14 attributes including the "pot" attribute. With no CI overlap between the groups for these variables and no zeros in the interval, it is appropriate to conclude that there is true difference in means between the two groups for these 14 attributes. Therefore, by employing both t-tests and bootstrap analysis, I have obtained a more comprehensive understanding of the mean differences and have reinforced the conclusion that the presence of chronic kidney disease (CKD) significantly impacts these variables compared to non-CKD cases.[Table 19]

The original correlation matrix and the permuted correlation matrix were compared further for evaluating the significance of correlation. The original correlation matrix represents the correlations between variables in the original dataset. The permuted correlation matrix, on the other hand, is calculated by permuting the values of the variables and then calculating the correlations. By comparing the two matrices, we can observe the differences in the correlation values. The permuted correlation matrix reflects the correlations that would be expected by chance if there were no underlying relationships between the variables. In other words, it represents the correlations that could arise due to random chance alone. The matrices suggests that there are significant differences between the original correlation matrix and the permuted correlation matrix, it suggests that the observed correlations in the original dataset are not solely due to random chance. Instead, they indicate the presence of meaningful relationships between the variables. To summarize, comparing the two matrices helps us assess the significance of the original correlations by evaluating them in the context of random permutations. The differences between the matrices indicate whether the observed correlations are likely to be meaningful or simply the result of random chance. The p-value represents the probability of observing the correlation coefficient (or a more extreme value) assuming there is no true correlation in the population. The significance level is used to determine whether the observed correlation is statistically significant or occurred by chance. If the p-value is less than the chosen significance level (commonly 0.05), it indicates that the correlation is statistically significant, and we reject the null hypothesis of no correlation. In the provided matrix, for most correlations, the p-values are very low (e.g., 0.001, 0.012, 0.009, etc.), indicating significant correlations. If the p-value is greater than the significance level, we fail to reject the null hypothesis, indicating that there is insufficient evidence to conclude a significant correlation. In the provided matrix, there are a few correlations with higher p-values (e.g., 0.062, 0.302, 0.169, etc.), indicating that these correlations are not statistically significant. To summarize, the matrix provides the p-values and their significance for each correlation analysis between different variables. By comparing

the p-values to the chosen significance level (0.05 in most cases), we can determine whether the correlations are statistically significant or not.

In the correlation matrix our only biggest concern was the correlation between two variables hemo and pcv with the correlation coefficient >0.8. Both the permuted matrix and the p values suggest that the correlation is significant and not solely based on chance. Thus, for the Classification and Regression only the attribute “hemo” was included in the analysis to avoid multicollinearity.[Table 20]

Based on the results from MC simulations, the mean accuracy of the LDA model across the Monte Carlo simulations is approximately 0.951, indicating that the model achieves a high level of overall correct classification. The mean precision is approximately 0.889, representing the proportion of true positives among all predicted positives. The mean sensitivity (recall) is approximately 0.995, indicating the proportion of true positives correctly identified. The mean F1 score, which combines precision and sensitivity, is approximately 0.939, reflecting the model's balance between precision and recall. The confusion matrix provided the summary of the classification results, on the original dataset showing the number of instances that were correctly or incorrectly classified by the model. True Positive (TP): There are 234 instances that belong to class 1 (chronic kidney disease, CKD) and were correctly classified as class 1. These are true positive predictions. True Negative (TN): There are 150 instances that belong to class 0 (not CKD, healthy) and were correctly classified as class 0. These are true negative predictions. False Positive (FP): There are 0 instances that actually belong to class 0 but were incorrectly classified as class 1. There are no false positive predictions. False Negative (FN): There are 16 instances that belong to class 1 but were incorrectly classified as class 0. These are false negative predictions. In this case, the model achieved a high accuracy, precision, sensitivity (recall), and F1 score, indicating that it performed well in classifying the instances into their respective classes. The absence of false positives (FP = 0) suggests that the model did not mistakenly classify any instances as CKD when they were not. However, there were a few false negatives (FN = 16) where instances of CKD were incorrectly classified as not CKD. The class probabilities obtained from the Monte Carlo simulations indicate the likelihood of each instance belonging to the positive class (1) or negative class (0). The average probabilities represent the average likelihood estimated by the LDA model across the simulations. The model achieves 100% correct classification on the original dataset, indicating a high level of accuracy in estimating class probabilities.

The permutation feature selection method, resulted in the ranked features with their respective importance scores. The results indicated that the "sg" feature has the highest importance score, followed by "al" and "sc." These features have a relatively higher impact on the model's performance compared to the other features. The attributes "hemo", "bu", "bgr", "bp" and "HTN" showed a moderate impact on the model's performance. The rest of the features had negligible importance scores.

Lastly, comparing the Logistic Regression models with all the attributes and the LR model with the attributes selected from the feature selection resulted in the observed difference in AUC-ROC between the full model and the original model is -0.000399999999999956. The

permutation test p-value is 0.98. Since the observed difference is negative and the p-value is greater than 0.05, there is no significant difference in predictive accuracy between the full model and the original model for classifying kidney disease.

Conclusion:

In this study, a dataset containing information related to kidney disease was examined. Initially, the dataset was evaluated for missing values, which were found in 1012 instances. To address the missing data, the "mice" package in R was employed, using the "pmm" imputation method. After imputing the missing values, a density plot was created to visualize the distribution of the imputed data. The imputed dataset was then further examined for missing values, and it was discovered that 234 missing values remained in the nominal columns. To resolve this issue, the mode of each nominal variable was computed, and the missing values were replaced with the corresponding mode values. By implementing these steps, all missing data in the dataset were successfully imputed. This allowed for a more comprehensive analysis, ensuring that the subsequent analysis and interpretations would not be biased due to missing values.

Based on the utilization of bootstrap resampling technique to estimate the population parameters helps in providing a reliable estimate of the population characteristics for each variable. The calculated median and mean offer insights into the central tendency, while the confidence intervals provide information about the precision and uncertainty associated with these estimates. The minimum and maximum values help identify the range within which the data are distributed.

Thirdly, bootstrap resampling and confidence interval estimation were performed on a dataset related to kidney disease. The results revealed the variability and uncertainty associated with various variables, including age, blood pressure, specific gravity, albumin, sugar, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, and red blood cell count. The calculated bootstrap confidence intervals allowed for a more comprehensive understanding of the potential range of values for each variable. These intervals provided insights into the variability of the population parameters, helping to assess the robustness of the sample estimates and the overall reliability of the data. In conclusion, the bootstrap analysis provided additional insights into the significance of mean differences between the CKD and non-CKD groups for various variables. It highlighted the variables that showed significant differences, indicating their potential importance in distinguishing between the two groups.

Correlation was performed on the dataset and permutation test was performed to test for the significance of the correlation. The original correlation matrix represents the correlations calculated from the actual data. The permuted correlation matrix represents the correlations calculated from the permuted (shuffled) datasets. The purpose of comparing the two matrices is to assess whether the observed correlations in the original data are significantly different from what would be expected by chance (random associations). By comparing the two matrices, you can identify if the correlations in the original data are statistically significant or if they could have occurred by chance. The p-value is a statistical measure that

quantifies the likelihood of obtaining a correlation as extreme or more extreme than the observed correlation, assuming the null hypothesis is true (i.e., assuming there is no true association between the variables). In the context of the permutation test, the p-value is determined by calculating the proportion of permuted correlation values that are more extreme than the corresponding values in the original correlation matrix. A p-value below a specified significance level (e.g., 0.05) indicates that the observed correlation is statistically significant, suggesting that it is unlikely to have occurred by chance alone. In this case, you can reject the null hypothesis and conclude that there is evidence of an association between the variables. On the other hand, a p-value above the significance level indicates that the observed correlation is not statistically significant, and there is insufficient evidence to reject the null hypothesis. This suggests that the observed correlation could have occurred by chance. In summary, comparing the two matrices helps assess whether the observed correlations in the original data are significantly different from random associations. The p-value, derived from this comparison, indicates the level of significance and determines whether the observed correlations are considered statistically significant or not. Based on the permutation test results, there is strong evidence to support significant correlations between the selected variables in the kidney disease dataset. The correlations observed in the original data are not due to random variation and are indicative of meaningful relationships. Correlation analysis helps us understand the strength and direction of the linear relationship between two variables. The correlation coefficient provides a measure of the strength and direction of the relationship.

The primary goal of performing MC simulation to assess the performance of the LDA model in classifying kidney disease. The use of Monte Carlo simulation allows to obtain a more comprehensive understanding of the model's performance by generating multiple random samples from the dataset and evaluating the model on each sample. Performing Monte Carlo simulations, helps to account for sampling variability. Randomly sampling the dataset with replacement generates different subsets, helps to assess the model's performance across a range of possible training datasets. This accounts for the inherent uncertainty in the data and provides a more robust estimate of the model's performance metrics. By conducting simulations and storing the performance metrics from each iteration, the mean and standard deviation of metrics such as accuracy, precision, sensitivity, and F1 score were calculated. These summary statistics provide an overall assessment of the model's performance and its consistency across different samples. To conclude, the LDA model demonstrates strong performance in classifying kidney disease. It achieves high accuracy, precision, sensitivity, and F1 score on average. These results suggest that the LDA model is effective in identifying patterns and discriminating between individuals with and without kidney disease.

The analysis of class probabilities demonstrates that the LDA model predicts the probabilities of class membership with high accuracy. All instances in the original dataset are correctly classified, as indicated by the "Correct" classification column. This suggests that the LDA model is able to accurately estimate the likelihood of kidney disease based on the given predictors. The analysis of the class probabilities predicted by the LDA model for the kidney disease dataset complements the performance evaluation using Monte Carlo Simulations by

focusing on the estimated probabilities rather than just the final classification. Analysing the class probabilities helps gain insights into the model's behaviour by understanding how the LDA model assigns likelihoods to different instances. It allows to explore the model's confidence in its predictions and provides a more granular understanding of its decision-making process. The percentage of correctly classified instances based on a threshold of 0.625 for probability provides an additional measure of the model's accuracy and verifies that the predicted probabilities align with the actual classification labels.

By using two different tasks, the evaluation of the LDA model from different angles were performed and more comprehensive insights were gathered. Task 1 focuses on overall performance metrics, while Task 2 provides a detailed analysis of class probabilities and the correctness of classification. Together, these tasks offer a more thorough assessment of the model's strengths, limitations, and behaviour.

The discrepancy between the analysis of class probabilities, which resulted in a 100% classification rate using a threshold of 0.625, and the presence of false negatives in the confusion matrix suggests that the LDA model's final classification decision incorporates factors beyond the predicted probabilities. These additional factors may influence the model's ability to correctly identify instances of chronic kidney disease (CKD) despite their predicted probabilities falling below the threshold. Therefore, while the analysis of class probabilities indicates a high accuracy based on the probabilities alone, the occurrence of false negatives in the confusion matrix highlights the impact of other considerations on the final classification outcome. The feature selection using permutation provides insights into the importance of each feature in the logistic regression model for classifying kidney disease. Based on the results, certain features such as "sg," "al," and "sc" contribute more significantly to the model's performance. Understanding the importance of these features can help in improving the interpretability and efficiency of the classification model.

Lastly, based on the AUC-ROC comparison and permutation testing, there is no significant evidence to suggest that the full model, which includes a subset of attributes selected through permutation testing, performs better than the original model in terms of predictive accuracy for classifying kidney disease. There is a possibility that the subset of attributes selected through permutation testing did not capture all the relevant information or interactions present in the original model, thus resulting in no difference between the two models. It's important to note that permutation testing evaluates feature importance based on the specific model and evaluation metric used. If the attributes chosen through permutation testing do not fully capture the underlying patterns or interactions in the data, it could result in a similar predictive performance between the full model and the original model.

Limitations:

Firstly, the imputation process assumes that the missing data are missing at random or missing completely at random. If this assumption is violated, the imputed values may not accurately reflect the true values, introducing potential bias. Additionally, the imputed

values are estimates based on observed data, leading to uncertainty in the analysis. The mode imputation method used for nominal variables may introduce bias if the most frequent category does not represent the true value. Furthermore, there is a risk of information loss during imputation, potentially impacting subsequent analyses. The findings from the imputed dataset may not generalize to the entire population, and biases may persist despite imputation. Interpretation of the imputed dataset requires caution, considering that imputed values are estimates and may not reflect the true values. Resource and time constraints can limit the extent of effective imputation. It is crucial to acknowledge these limitations when interpreting the results and consider alternative approaches to validate and address potential biases in the imputed data.

The limitation of using bootstrap resampling to estimate population parameters is that it assumes the resampled data is representative of the underlying population. While bootstrap resampling allows us to estimate the variability and uncertainty of summary statistics, it relies on the assumption that the original dataset is a good representation of the population. If the original dataset is biased or contains sampling errors, the bootstrap resampling results may also be biased or inaccurate. Therefore, it is crucial to ensure that the original dataset is representative and accurately reflects the population of interest before applying bootstrap resampling.

It is important to acknowledge the limitations of the permutation test and this analysis. Permutation tests assume exchangeability of variables under the null hypothesis, which might not hold in all cases. Additionally, the number of permutations chosen may impact the accuracy and stability of the results. Other limitations include the specific dataset used and potential biases within it.

The limitation of the evaluation of classification analysis is that it assumes the LDA model is appropriate for the given dataset and that the predictors are linearly related to the response variable. If these assumptions are violated, the LDA model may not provide accurate results. Additionally, the evaluation is performed on the same dataset used for training, which may overestimate the model's performance. It would be beneficial to evaluate the model on an independent test dataset to assess its generalizability.

Additionally, the class probabilities assumes a binary classification problem, where the positive class represents kidney disease and the negative class represents the absence of kidney disease. If there are other classes or subtypes within kidney disease, this analysis may not capture the nuances of those subcategories.

The feature selection method assumes that a linear relationship exist between the features and the target variable. If the relationship is nonlinear or complex, the importance scores may not accurately reflect the true impact of the features. Additionally, the feature selection process is specific to the logistic regression model used in this study and may not be directly applicable to other classification algorithms.

Performing feature selection through permutation is only considering AUC-ROC as the evaluation metric in this analysis. Other evaluation metrics such as accuracy, precision, recall, F1 score, and log-likelihood or deviance could provide additional insights into the

model's performance. Furthermore, the number of permutations (1000) may influence the stability of the p-value estimation. Furthermore, there is a chance of the omission of certain attributes or interactions through feature selection using permutation could limit the ability of the reduced model to differentiate between kidney disease and non-kidney disease cases. Thirdly, the limitation of using only 7 attributes selected through permutation testing is that it may not capture the full complexity of the underlying data. By limiting the model to a subset of attributes, there is a possibility of omitting important variables that could contribute to the predictive accuracy. Also, the feature selection based on permutation testing may result in overfitting to the specific dataset used. The selected attributes may perform well on the current dataset, but their performance could vary when applied to different datasets or real-world scenarios. Therefore, the generalizability of the model's performance to new data or populations may be limited.

Lastly, both the logistic regression and classification were performed on a single dataset without utilizing separate training and test datasets. Using a single dataset for both model training and evaluation can lead to overfitting, where the model learns the specific patterns and noise present in the training data rather than generalizing well to unseen data. As a result, the reported performance measures, such as AUC-ROC, may be overly optimistic and not reflective of the model's true performance on new, unseen data.

Recommendation:

Based on the analysis conducted, it is wise to proceed with the imputed dataset for further analysis. Imputing the missing data using the "pmm" method and handling the missing values in the nominal columns by replacing them with the mode values have provided a more complete dataset. However, it is important to note that the imputation process introduces uncertainty, as imputed values are estimates based on observed data. Therefore, it is advisable to interpret the results with caution and acknowledge the potential limitations associated with imputed data. Additionally, it is recommended to perform appropriate statistical analyses and modelling techniques on the imputed dataset to explore relationships, identify patterns, and draw meaningful conclusions related to kidney disease.

Given the robustness and flexibility of bootstrap resampling, one recommendation to utilize this methodology for calculating summary statistics in similar data analysis tasks. By capturing the inherent variability and uncertainty present in the dataset, bootstrap resampling provides more reliable and informative summary statistics compared to traditional approaches.

Based on the analysis conducted, it can be recommended to further investigate and prioritize the variables that exhibited significant differences (age, bp, sg, al, su, bgr, bu, sc, sod, and hemo) as potential markers for identifying CKD cases. The variable "pot" may not be as informative in this regard and could be given less emphasis in future analyses or diagnostic models. Additionally, it is wise to consider the bootstrap results, including other simulation techniques such as permutation, to further emphasis if there is true difference between the means when interpreting and analyzing the dataset related to kidney disease.

The bootstrap methodology to test for difference in means provides a valuable tool for quantifying uncertainty and variability, enhancing the reliability of statistical inferences and supporting robust decision-making in real-life scenarios.

Considering the lack of statistically significant correlations, it may be necessary to explore other variables or employ alternative analytical techniques to uncover potential relationships or predictors for kidney disease. Additionally, it is important to consider domain knowledge, biological plausibility, and other relevant factors in the analysis and interpretation of the data.

To further validate the performance of the LDA model and ensure its generalizability, it is a necessity to perform cross-validation or hold-out validation using an independent dataset. This will provide a more reliable estimate of the model's performance on unseen data. Additionally, considering the assumption of linear relationship in LDA, exploring other classification algorithms that can capture non-linear relationships may be beneficial.

To gain a better understanding of the LDA model's performance and assess its robustness, a further analysis using additional evaluation metrics, such as receiver operating characteristic (ROC) curve analysis needs to be conducted, to evaluate the model's ability to discriminate between the two classes. Additionally, investigating the specific instances where the model made incorrect classifications can provide insights into the model's limitations and potential areas for improvement.

To further validate the findings and ensure the generalizability of the feature importance, it may be beneficial to apply this feature selection method to other classification algorithms and compare the results. Additionally, conducting cross-validation or using an independent test dataset can provide a more robust evaluation of feature importance. Exploring alternative feature selection techniques, such as recursive feature elimination or random forests, can also complement the findings and provide a comprehensive understanding of feature importance. It is important to note that the process of feature selection is dependent on the specific dataset and the goals of the analysis. While permutation testing can provide valuable insights, it is crucial to consider other techniques, such as cross-validation or information criteria, to ensure robust feature selection and model evaluation. Exploring alternative feature selection methods like recursive feature elimination, random forests, or LASSO regression can provide different perspectives on feature importance. Additionally, considering factors like data quality, sample size, or unmeasured confounders can help understand why the reduced model did not show a significant difference in predictive accuracy compared to the original model. In summary, the limitations of the attributes selected through permutation testing may have contributed to the lack of difference between the models. Exploring alternative techniques and conducting further analysis can help identify a more informative subset of attributes for predicting kidney disease.

To mitigate the limitation of the reported performance measures, such as AUC-ROC, may be overly optimistic and not reflective of the model's true performance on new, unseen data, splitting the dataset into training and test sets may improve the performance. The training

set is used to build the models, while the test set is used to evaluate the models' performance on unseen data. By evaluating the models on an independent test set, it provides a more realistic assessment of their generalization ability and predictive accuracy. Using cross-validation techniques, such as k-fold cross-validation, can also help in obtaining a more robust evaluation by repeatedly splitting the data into training and test sets and averaging the performance across multiple iterations. By incorporating separate training and test datasets or employing cross-validation, it helps to ensure that the reported performance of the models is more reliable and representative of their actual predictive performance on new, unseen data.

References:

Singh, A. K., Farag, Y. M. K., Mittal, B. V., Subramanian, K., Reddy, S. R., Acharya, V. N., & Almeida, A. F. (2013). Epidemiology and risk factors of chronic kidney disease in India: Results from the SEEK (Screening and Early Evaluation of Kidney Disease) study. BMC Nephrology, 14(1), 114. doi:10.1186/1471-2369-14-114

Vink, G., Frank, L. E., Pannekoek, J., & van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. Statistical Methods in Medical Research, 23(5), 417-434. doi:10.1177/0962280213503432

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. Journal of the American Statistical Association, 89(426), 446-447. doi:10.1080/01621459.1994.10476470

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. Austral Ecology, 26(1), 32-46. doi:10.1111/j.1442-9993.2001.01070.pp.x

Davison, A. C., & Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge University Press.

Tables and Graphs:

id	**age**	**bp**	**sg**	**al**
Min. : 0.00	Min. : 2.00	Min. : 50.00	Min. :1.005	Min. :0.000
1st Qu.: 99.75	1st Qu.:42.00	1st Qu.: 70.00	1st Qu.:1.015	1st Qu.:0.000
Median :199.50	Median :55.00	Median : 80.00	Median :1.020	Median :0.000
Mean :199.50	Mean :51.71	Mean : 76.33	Mean :1.018	Mean :1.095
3rd Qu.:299.25	3rd Qu.:65.00	3rd Qu.: 80.00	3rd Qu.:1.020	3rd Qu.:2.000
Max. :399.00	Max. :90.00	Max. :180.00	Max. :1.025	Max. :5.000
su	**rbc**	**pc**	**pcc**	**ba**
Min. :0.00	Length:400	Length:400	Length:400	Length:400
1st Qu.:0.00	Class :character	Class :character	Class :character	Class :character
Median :0.00	Mode :character	Mode :character	Mode :character	Mode :character
Mean :0.46				
3rd Qu.:0.00				
Max. :5.00				

bgr	bu	sc	sod	pot
Min. : 22.0	Min. : 1.50	Min. : 0.400	Min. : 4.5	Min. : 2.500
1st Qu.: 99.0	1st Qu.: 27.00	1st Qu.: 0.900	1st Qu.:135.0	1st Qu.: 3.800
Median :121.0	Median : 42.00	Median : 1.200	Median :138.0	Median : 4.400
Mean :149.5	Mean : 57.32	Mean : 3.016	Mean :136.9	Mean : 4.587
3rd Qu.:169.0	3rd Qu.: 66.00	3rd Qu.: 2.800	3rd Qu.:142.0	3rd Qu.: 4.900
Max. :490.0	Max. :391.00	Max. :76.000	Max. :163.0	Max. :47.000
hemo	**pcv**	**wc**	**rc**	**htn**
Min. : 3.10	Min. : 9.0	Min. : 2200	Min. :2.100	Length:400
1st Qu.:10.40	1st Qu.:32.0	1st Qu.: 6500	1st Qu.:3.900	Class :character
Median :12.70	Median :40.0	Median : 7950	Median :4.600	Mode :character
Mean :12.55	Mean :38.5	Mean : 8458	Mean :4.589	
3rd Qu.:15.00	3rd Qu.:45.0	3rd Qu.: 9800	3rd Qu.:5.225	
Max. :17.80	Max. :54.0	Max. :26400	Max. :8.000	
dm	**cad**	**appet**	**pe**	**ane**
Length:400	Length:400	Length:400	Length:400	Length:400
Class :character	Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character	Mode :character
classification	**classify**	**RBC**	**PC**	**PCC**
Length:400	Min. :0.000	Min. :0.0000	Min. :0.00	Min. :0.000
Class :character	1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:0.00	1st Qu.:0.000
Mode :character	Median :1.000	Median :0.0000	Median :0.00	Median :0.000
	Mean :0.625	Mean :0.1175	Mean :0.19	Mean :0.105
	3rd Qu.:1.000	3rd Qu.:0.0000	3rd Qu.:0.00	3rd Qu.:0.000
	Max. :1.000	Max. :1.0000	Max. :1.00	Max. :1.000
HTN	**DM**	**APPET**	**PE**	**ANE**
Min. :0.0000	Min. :0.0000	Min. :0.000	Min. :0.00	Min. :0.00
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:1.000	1st Qu.:0.00	1st Qu.:0.00
Median :0.0000	Median :0.0000	Median :1.000	Median :0.00	Median :0.00
Mean :0.3675	Mean :0.3425	Mean :0.795	Mean :0.19	Mean :0.15
3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.000	3rd Qu.:0.00	3rd Qu.:0.00
Max. :1.0000	Max. :1.0000	Max. :1.000	Max. :1.00	Max. :1.00
CAD	**AGE**	**BA**		
Min. :0.000	Min. :0.0000	Min. :0.000		
1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:0.000		
Median :0.000	Median :1.0000	Median :0.000		
Mean :0.085	Mean :0.5475	Mean :0.055		
3rd Qu.:0.000	3rd Qu.:1.0000	3rd Qu.:0.000		
Max. :1.000	Max. :1.0000	Max. :1.000		

Table 2: Summary Statistics

Variables	Description	Levels of Measurement
Age	Age in years	Continuous
Blood Pressure	Diastolic blood pressure in mm/Hg	Continuous
Specific Gravity	Urine specific gravity	Discrete
Albumin	Presence of albumin in urine	Discrete/Categorical
Sugar	Presence of sugar in urine	Discrete/Categorical
Red Blood Cells	Presence of abnormal red blood cells in urine	Discrete/Categorical
Pus Cell	Presence of abnormal pus cells in urine	Discrete/Categorical
Pus Cell clumps	Presence of pus cell clumps in urine	Discrete/Categorical
Bacteria	Presence of bacteria in urine	Discrete/Categorical
Blood Glucose Random	Random blood glucose level in mg/dL	Continuous
Blood Urea	Blood urea level in mg/dL	Continuous
Serum Creatinine	Serum creatinine level in mg/dL	Continuous
Sodium	Sodium level in mEq/L	Continuous
Potassium	Potassium level in mEq/L	Continuous
Hemoglobin	Hemoglobin level in gms	Continuous
Packed Cell Volume	Volume percentage of packed red blood cells	Continuous
White Blood Cell Count	White blood cell count per cubic millimeter	Continuous
Red Blood Cell Count	Red blood cell count per cubic millimeter	Continuous
Hypertension	Presence of hypertension	Categorical
Diabetes Mellitus	Presence of diabetes mellitus	Categorical
Coronary Artery Disease	Presence of coronary artery disease	Categorical
Appetite	Appetite status	Categorical
Pedal Edema	Presence of pedal edema	Categorical
Anemia	Presence of anemia	Categorical
classification	Presence of chronic kidney disease (CKD)	Categorical

Table 3: Description of Variables

Missing Values in the Original Dataset

	age	bp	sg	al	
0	9	12	47	46	

sc	sod	pot	hemo	pcv	
17	87	88	52	71	
su	rbc	pc	pcc	ba	
49	152	65	4	4	
wc	rc	htn	dm	cad	
106	131	2	2	2	
bgr	bu	ane	classification	appet	pe
44	19	1	0	1	1

Table 4: Missing Data Before Imputation

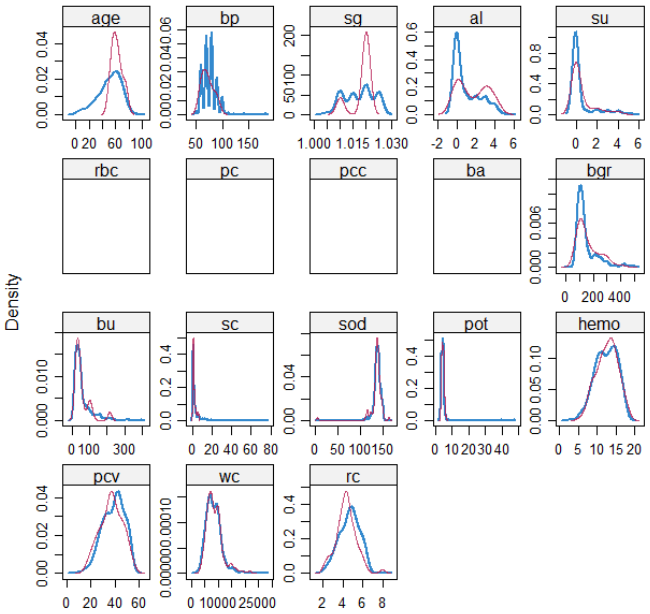

Table 5: Density plot of the Imputed Data using "mice" package

---

Missing Values After Imputation

id	age	bp	sg	ba
0	0	0	0	0
sc	sod	pot	hemo	cad
0	0	0	0	0
al	su	rbc	pc	pcc
0	0	0	0	0
pcv	wc	rc	htn	dm
0	0	0	0	0
bgr	bu	ane	classification	
0	0	0	0	
appet	pe			
0	0			

Table 6: Missing Data After Imputation

Variable	Category Meaning
classify	CKD (1), Non-CKD (0)
RBC	Normal (0), Abnormal (1)
PC	Normal (0), Abnormal (1)
PCC	Not Present (0), Present (1)
BA	Not Present (0), Present (1)
HTN	No (0), Yes (1)
DM	No (0), Yes (1)
APPET	Good (1), Poor (0)
PE	No (0), Yes (1)
ANE	No (0), Yes (1)
CAD	No (0), Yes (1)
AGE	Age <= 52 (0), Age > 52 (1)

Table 7: Extra Attributes used in the analysis

	variable	min	lower_ci	median	mean	upper_ci	max
1	age	2	50.1080716	55	51.7014275	53.3462297	90
2	bp	50	75.0256351	80	76.3429	77.7743649	180
3	sg	1.005	1.0169625	1.02	1.0174895	1.0180372	1.025
4	al	0	0.960127	0	1.0939525	1.2324365	5
5	su	0	0.3525635	0	0.4599575	0.5699365	5
6	bgr	22	141.5727541	121	149.44143	157.8473095	490
7	bu	1.5	52.6684654	42	57.2400845	62.3422546	391
8	sc	0.4	2.5097234	1.2	3.0108829	3.6626166	76
9	sod	4.5	135.5490358	138	136.8378675	137.9071824	163
10	pot	2.5	4.3662691	4.4	4.582286	4.8929047	47
11	hemo	3.1	12.2625953	12.7	12.5394283	12.8072309	17.8
12	pcv	9	37.625127	40	38.5072725	39.4148095	54
13	wc	2200	8169.440538	7950	8459.57925	8752.75	26400
14	rc	2.1	4.4872627	4.6	4.5889065	4.68325	8

Table 8: Population Estimation using Bootstrap

Variable	t-value	p-value	Significant Difference
age	4.862223	1.78E-06	Yes
bp	6.744965	5.44E-11	Yes
sg	-20.3445	1.57E-63	Yes
al	19.81559	4.23E-53	Yes
su	8.870203	1.45E-16	Yes
bgr	11.38794	5.62E-25	Yes
bu	9.936435	3.26E-20	Yes
sc	7.826171	1.40E-13	Yes
sod	-8.14926	7.40E-15	Yes

pot	1.765794	0.0785549	No
hemo	-22.8853	2.74E-74	Yes
pcv	-21.1063	1.07E-66	Yes
wc	4.639764	4.75E-06	Yes
rc	-15.9424	1.36E-44	Yes

Table 9: T-test

Variable	CI for CKD	CI for Non-CKD	Overlap	If Overlap = 0	Significant Difference
Age	52.56 - 56.95	43.92 - 48.87	No	Significant	Yes
BP	77.47 - 81.20	69.67 - 72.60	No	Significant	Yes
SG	1.01 - 1.02	1.02 - 1.03	No	Significant	Yes
AL	1.58 - 1.92	1.02 - 1.03	No	Significant	Yes
SU	0.59 - 0.91	1.02 - 1.03	No	Significant	Yes
BGR	163.96 - 186.24	103.92 - 109.80	No	Significant	Yes
BU	64.27 - 79.43	31.48 - 35.72	No	Significant	Yes
SC	3.61 - 5.45	0.85 - 0.97	No	Significant	Yes
Sod	131.5448 - 135.3138	140.905 - 142.48	No	Significant	Yes
Pot	4.433133 - 5.478502	4.243333 - 4.424667	No	Significant	Yes
Hemo	10.69884 - 11.2832	14.98941 - 15.38068	No	Significant	Yes
Pcv	32.77718 - 34.68834	45.66055 - 47.0052	No	Significant	Yes
Rc	3.994444 - 4.250089	5.272486 - 5.46247	No	Significant	Yes
Wc	8516.724 - 9366.045	7411.223 - 8006.959	No	Significant	Yes

Table 10: Bootstrapped CI and the resulting Significance

	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc
age	1	0.145	0.172	0.127	0.213	0.235	0.193	0.136	0.092	0.052	0.213	0.244	0.103	0.276
bp	0.145	1	0.168	0.128	0.212	0.171	0.193	0.144	0.048	0.062	0.266	0.299	0.08	0.201
sg	0.172	0.168	1	0.407	0.249	0.314	0.28	0.22	0.256	0.054	0.493	0.477	0.199	0.425
al	0.127	0.128	0.407	1	0.245	0.354	0.442	0.3	0.303	0.099	0.573	0.554	0.232	0.436
su	0.213	0.212	0.249	0.245	1	0.72	0.144	0.221	0.158	0.174	0.179	0.191	0.139	0.18

bgr	0.235	0.171	0.314	0.354	0.72	1	0.133	0.084	0.181	0.047	0.293	0.285	0.137	0.229
bu	0.193	0.193	0.28	0.442	0.144	0.133	1	0.591	0.21	0.333	0.584	0.588	0.051	0.552
sc	0.136	0.144	0.22	0.3	0.221	0.084	0.591	1	0.704	0.201	0.395	0.419	0.001	0.385
sod	0.092	0.048	0.256	0.303	0.158	0.181	0.21	0.704	1	0.062	0.306	0.324	0.006	0.259
pot	0.052	0.062	0.054	0.099	0.174	0.047	0.333	0.201	0.062	1	0.147	0.176	0.106	0.179
hemo	0.213	0.266	0.493	0.573	0.179	0.293	0.584	0.395	0.306	0.147	1	0.899	0.136	0.792
pcv	0.244	0.299	0.477	0.554	0.191	0.285	0.588	0.419	0.324	0.176	0.899	1	0.153	0.784
wc	0.103	0.08	0.199	0.232	0.139	0.137	0.051	0.001	0.006	0.106	0.136	0.153	1	0.092
rc	0.276	0.201	0.425	0.436	0.18	0.229	0.552	0.385	0.259	0.179	0.792	0.784	0.092	1

Table 11: Correlation Matrix

	age	**bp**	**sg**	**al**	**su**	**bgr**	**bu**	**sc**	**sod**	**pot**	**hemo**	**pcv**	**wc**	**rc**
age	1	0	0.001	0.001	0	-0.001	0.001	0	0.001	0.001	-0.001	0	0.001	0.003
bp	0	1	0.002	0.002	0.003	0.001	0.001	0.001	0	0.002	0.001	0	0.002	0.001
sg	0.001	0.002	1	0.003	0.001	0.001	0.002	0.002	0	0.001	0	0.001	0.001	0.003
al	0.001	0.002	0.003	1	0.002	0.002	0.001	0	0	0	0.001	0.001	0.001	0
su	0	0.003	0.001	0.002	1	0	0.001	0	0.001	0.001	0.001	0.001	0.002	0.002
bgr	0.001	0.001	0.001	0.002	0	1	0.001	0.001	0	0.001	0.001	0.001	0.004	0.003
bu	0.001	0.001	0.002	0.001	0.001	0.001	1	0.001	0.001	0	0.001	0.004	0.002	0
sc	0	0.001	0.002	0	0	0.001	0.001	1	0.003	0	0.003	0.001	0	0.001
sod	0.001	0	0	0	0.001	0	0.001	0.003	1	0.002	0	0.002	0.002	0.001
pot	0.001	0.002	0.001	0	0.001	0.001	0	0	0.002	1	0.001	0.002	0	0.003
hemo	0.001	0.001	0	0.001	0.001	0.001	0.001	0.003	0	0.001	1	0.001	0.003	0.001
pcv	0	0	0.001	0.001	0.001	0.001	0.004	0.001	0.002	0.002	0.001	1	0.002	0.002
wc	0.001	0.002	0.001	0.001	0.002	0.004	0.002	0	0.002	0	0.003	0.002	1	0.001

rc	0.003	0.001	0.003	0	0.002	0.003	0	0.001	0.001	0.003	0.001	0.002	0.001	1

Table 12: Permuted Correlation Matrix

P- Values	Significance
p-value for correlation between age and bp : 0.001	**Yes**
p-value for correlation between age and sg : 0.001	**Yes**
p-value for correlation between age and al : 0.012	**Yes**
p-value for correlation between age and su : 0	**Yes**
p-value for correlation between age and bgr : 0	**Yes**
p-value for correlation between age and bu : 0	**Yes**
p-value for correlation between age and sc : 0.009	**Yes**
p-value for correlation between age and sod : 0.062	**No**
p-value for correlation between age and pot : 0.302	**No**
p-value for correlation between age and hemo : 0	**Yes**
p-value for correlation between age and pcv : 0	**Yes**
p-value for correlation between age and wc : 0.039	**Yes**
p-value for correlation between age and rc : 0	**Yes**
p-value for correlation between bp and age : 0.001	**Yes**
p-value for correlation between bp and sg : 0	**Yes**
p-value for correlation between bp and al : 0.009	**Yes**
p-value for correlation between bp and su : 0	**Yes**
p-value for correlation between bp and bgr : 0	**Yes**
p-value for correlation between bp and bu : 0	**Yes**

Table 13: Sample P-values to assess the significance of correlation

	Predicted Labels	
Actual Labels	**0**	**1**
0	**150**	**0**
1	**16**	**234**

Table 14: Confusion Matrix

"Mean Accuracy: 0.95145925" "Mean Precision: 0.88870634872856" "Mean Sensitivity (Recall): 0.99555" "Mean F1 Score: 0.939028655054618"

Table 15: Evaluation Metrics

Percentage of Correctly Classified Instances: 100 %

Table 16: Percentage of correctly classified instances

	Attribute	Importance
3	sg	0.056
4	al	0.03
12	sc	0.022
15	hemo	0.014
11	bu	0.008
10	bgr	0.007
2	bp	0.003
18	HTN	0.002
1	age	0
17	rc	0
5	su	0
6	RBC	0
7	PC	0
8	PCC	0
9	BA	0
13	sod	0
14	pot	0
16	wc	0
19	DM	0
20	APPET	0
21	PE	0

22	ANE	0
23	CAD	0

Table 17: Feature Selection using Permutation

Observed Difference in AUC-ROC: -0.000399999999999956
Permutation Test p-value: 0.98

Table 18: Model Comparison

Variable	Bootstrap Result	T-Test Result	Bootstrap Advantage
age	Significant	Significant	Similar
bp	Significant	Significant	Similar
sg	Significant	Significant	Similar
al	Significant	Significant	Similar
su	Significant	Significant	Similar
bgr	Significant	Significant	Similar
bu	Significant	Significant	Similar
sc	Significant	Significant	Similar
sod	Significant	Significant	Similar
pot	Significant	Not Significant	Different Results
hemo	Significant	Significant	Similar

Table 19: Significance of Ttest vs Bootstrap CI.

Permutation test result of hemo and pcv	**-0.001**
Correlation between hemo and pcv	**0.899**
p-value for correlation between hemo and pcv : 0	**Yes**

Table 20: Multicollinearity between "hemo" and "pcv"