Comprehensive Report on Chronic Kidney Disease (CKD) Analysis

Data Science II

COSC 4337

Submitted to

Dr. Ricardo Vilalta

Submitted by

Joseph Irving (1766731)

Daniel Emami (1390157)

Ryan Nguyen (1897135)

Executive Summary

Chronic Kidney Disease (CKD) is a significant global health issue that impacts millions of individuals all over the world by progressively impairing kidney function. However, through early detection and intervention, this can substantially delay the progression toward severe stages of the disease, potentially avoiding the need for invasive treatments such as a kidney transplantation or dialysis. This report discusses our analytical approach to utilizing clinical health data to discover early predictors of CKD. Our main objectives were to identify early indicators of CKD from a vast array of health data points which in turn empowers healthcare professionals with having the reliable data to make important medical decisions with their patients. Through careful data handling and rigorous analysis, we identified several key health indicators strongly predictive of CKD. These findings aim to provide actionable insights that can significantly enhance the efficacy of screening programs and patient education regarding kidney health.

Introduction

Chronic Kidney Disease affects an estimated 10% of the global population and is a major challenge due to its asymptomatic nature in the early stages and complex management requirements in later stages. The disease usually progresses silently over many years and often remains undetected until it reaches an advanced stage, making management and treatment increasingly difficult. The need for effective early detection tools is critical to change the disease's trajectory for many patients. This project was initiated to explore how data analysis techniques could identify potential predictors of CKD from extensive health data. By understanding these predictors, healthcare providers can implement earlier interventions and potentially improve patient outcomes, thereby reducing the burdens on healthcare systems worldwide.

Data Collection and Preparation

The data collection and preparation phase were foundational to our analysis, involving meticulous data handling to ensure the accuracy and reliability of our findings. Data from the CKD dataset spans from a two-month period in India with features. Our dataset was extensive,

containing a diverse range of patient records including comprehensive medical histories, laboratory test results, and demographic information. Key steps in our data preparation included:

- **Data Cleaning:** The dataset to correct any inaccuracies such as duplicates, incorrect entries, and outliers that could potentially skew our analysis results.
- **Handling Missing Information:** We tackled the challenge of missing data robustly by using advanced imputation methods that estimate missing values based on available data. This approach helped us maintain the dataset's completeness and reliability without introducing significant biases.
- **Standardizing Data:** We standardized the measurements across various health indicators to ensure that all data adhered to a uniform scale. This standardization was crucial for later stages of analysis, allowing for accurate comparisons and assessments across different types of health data.

Analysis Methodology

Our analysis methodology was designed to be comprehensive yet understandable for non-technical stakeholders, focusing on practical outcomes that could directly influence patient care:

Identifying Key Health Indicators: We used straightforward statistical techniques to identify patterns and correlations within the data, focusing on those that recurred frequently among individuals diagnosed with CKD.

Developing Predictive Insights: The essence of our analysis was to determine which factors most significantly indicate the risk of developing CKD. This involved an in-depth comparison of the health profiles of individuals with CKD against those without the disease, identifying the most impactful differentiators.

Key Findings

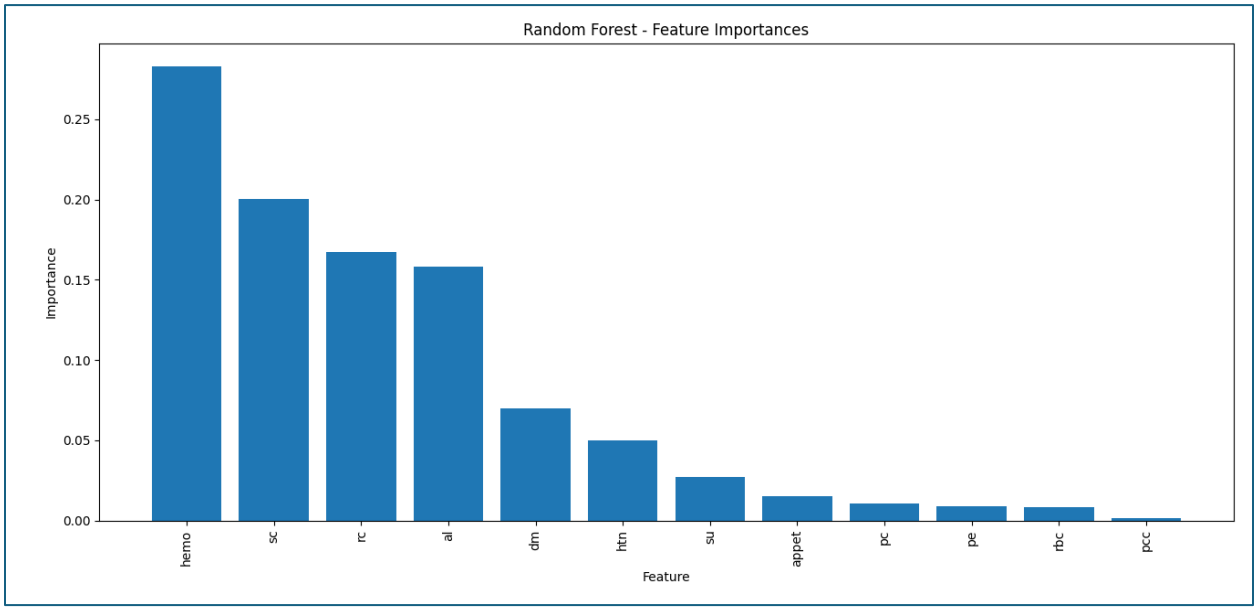
Our investigative process yielded significant insights into the health indicators that are most predictive of CKD:

High Blood Pressure and Elevated Blood Glucose Levels: These were among the most significant predictors found. Our analysis highlighted the strong link between unmanaged hypertension and diabetes with the development of CKD, suggesting that managing these conditions may be key to preventing or delaying the onset of kidney disease.

Biochemical Indicators: We found that serum creatinine and hemoglobin levels are crucial in predicting CKD. Specifically, elevated creatinine levels often indicate reduced kidney function, while low hemoglobin can point to anemia related to kidney disease.

Visual Representation of Data

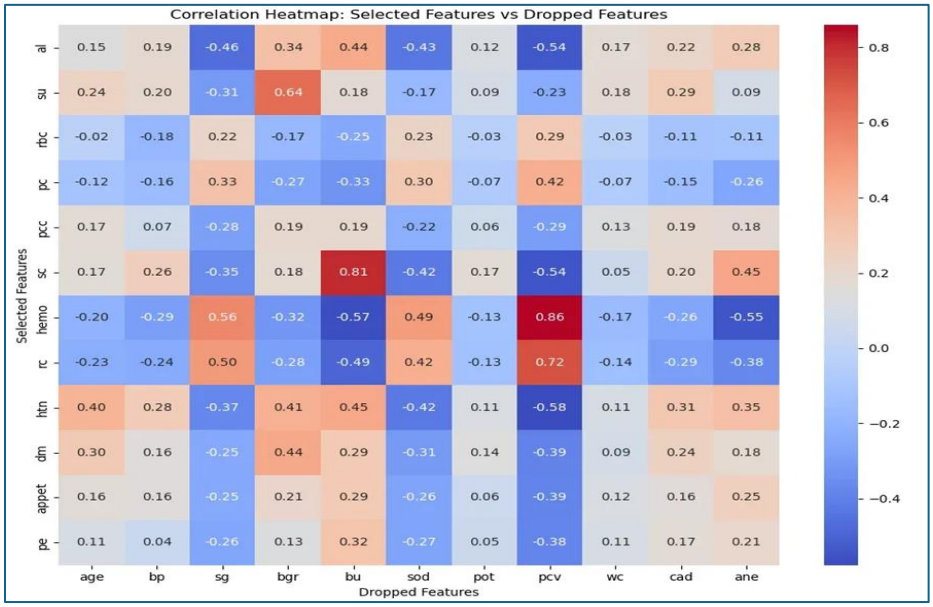
Our use of visual data representations was strategic, aimed at making the data insights accessible and understandable:



Bar Chart of Health Indicators: The provided bar chart represents the results from a data analysis model known as Random Forest, which helps us understand the most influential health factors in predicting CKD. In simple terms, the chart ranks various health indicators based on their importance or relevance in determining the presence of CKD. Hemoglobin levels, indicated by 'hemo' on the chart, are shown as the most significant indicator, suggesting that they are a crucial factor in CKD detection. Serum creatinine, labeled as 'sc', is the second most important, underscoring its role as a key indicator of kidney function. Other factors like red blood cell count

and albumin levels also play significant roles but to a lesser extent. This visual tool helps us quickly identify which health parameters are most critical in assessing the risk or presence of kidney disease, enabling more focused and effective medical examinations and interventions.

Trend Lines Over Time: These graphs depict how variations in key indicators correlate with the progression of CKD, visually narrating the impact of each factor over time.

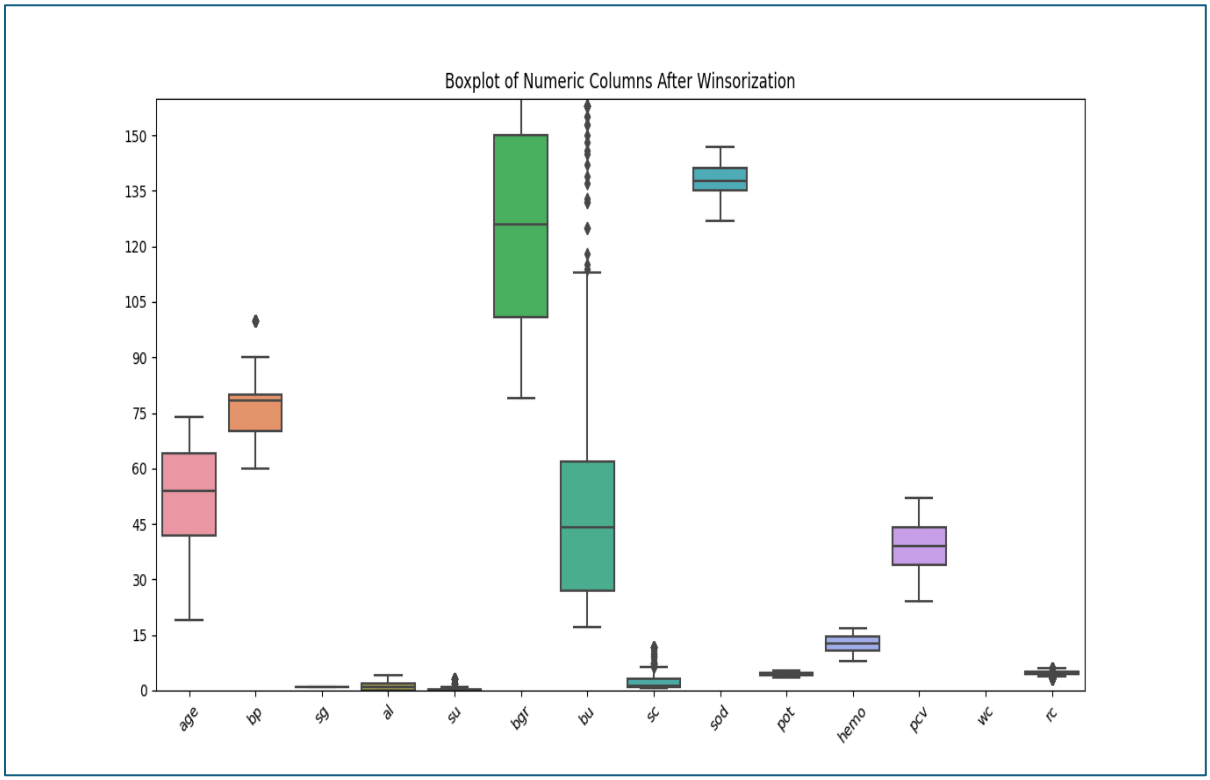


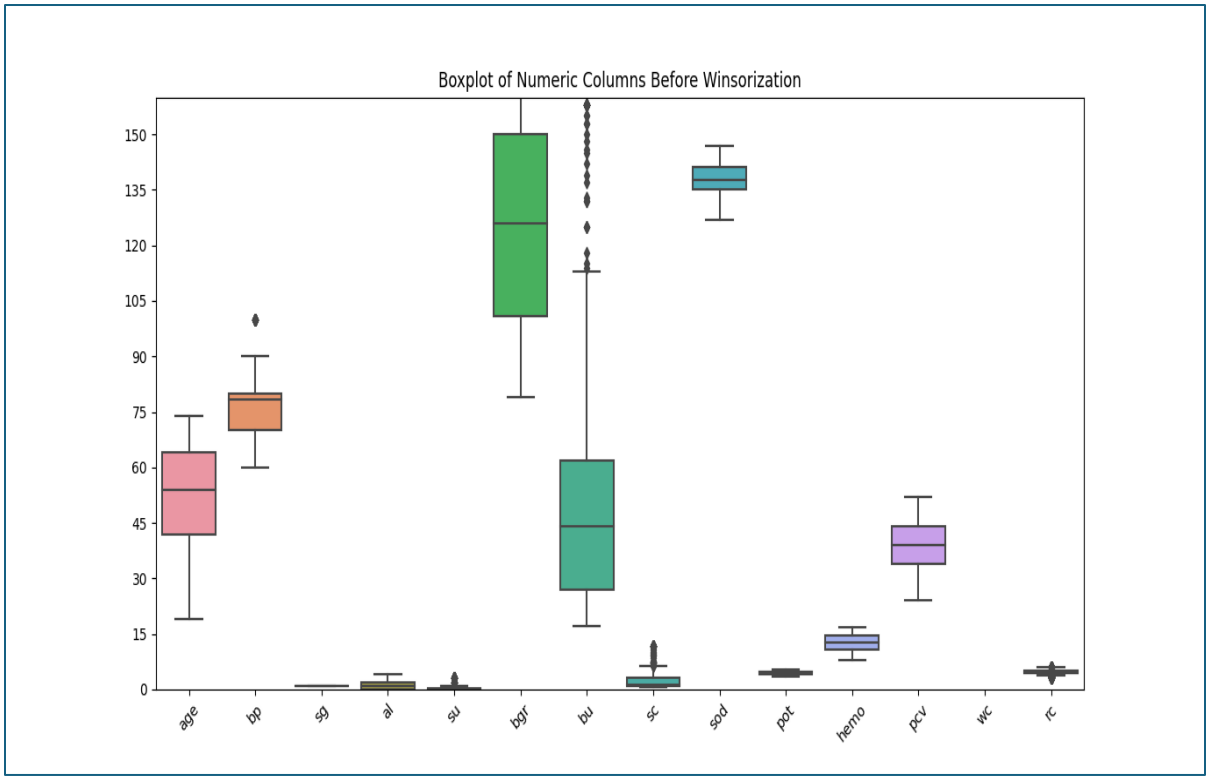
Part of our analysis included finding the most relevant features associated with CKD that our model could use to predict whether a patient had this condition. We started by preprocessing the dataset, which included making decisions about how to handle things like missing data or outliers (values well outside the norm). Then, we ran the data through a technique called Recursive Feature Elimination, which determines the most informative variables and eliminates the less important ones. This allows us to focus on the key variables that contribute most to predicting CKD.

From a healthcare professional's perspective, it’s crucial to understand that the features eliminated from our model, while less important for prediction, still hold clinical relevance when evaluating a patient for CKD. The heatmap provided above illustrates correlations between selected features and dropped features. The darker red a square is, the more positively correlated those features are (as one goes up, so does the other), and the darker blue a square is the more negatively correlated they are (as one goes up, the other goes down). Another way

to interpret this is that anything greater than 0.7 or less than -0.7 indicates a strong correlation, with 1 or -1 indicating the strongest correlation. This provides valuable insight into the interconnectedness of these values.

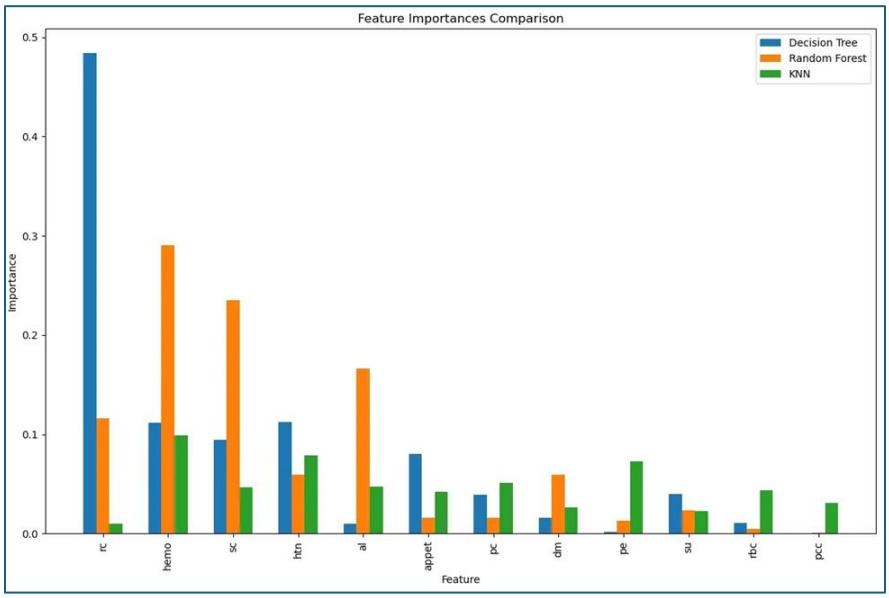
For example, we can see very strong correlations between ‘pcv’ (packed cell volume) and ‘hemo’ (hemoglobin) as well as ‘wc’ (white blood cell count) and ‘rc’ (red blood cell count). Looking at the feature importance chart provided above, we can see that ‘hemo’ and ‘rc’ are both significant features for our models and it is reasonable to believe that these dropped features still hold valuable information as they relate to CKD. It is possible these features were dropped to reduce redundant data and improve performance. The key takeaway here is that these features should not be disregarded completely, and the healthcare professional should incorporate their understanding of these features along with the model’s predictions to gain a more comprehensive view of the patient's condition to improve decision-making related to their care.





Before Adjustments: The "Boxplot of Numeric Columns Before Winsorization" shows how patient health indicators like kidney function (serum creatinine), waste in the blood (blood urea), and potassium levels are distributed before we make any adjustments. In this chart, you can see several points that stand far outside the general cluster of data, representing unusually high or low values. While these extreme values can sometimes highlight severe or unique cases of kidney disease, they can also make it harder to understand what's typical or common among most patients. This wide variation can skew our perspective, making the average seem worse or better than it actually is.

After Adjustments: The "Boxplot of Numeric Columns After Winsorization" visual shows these same health indicators after we've adjusted for these extreme values. By trimming these extremes, we bring them closer to a range that represents the majority of CKD patients more accurately. This makes the overall picture more compact and focused, reducing the distraction caused by rare, extreme cases. The charts now show fewer outlying points, making it easier to see and understand the common patterns and trends among most patients with CKD. This comparative visualization before and after Winsorization serves a crucial function in our data preprocessing strategy.



The feature importance chart presented above is not just a theoretical concept, but a practical tool that can significantly aid healthcare professionals in their daily tasks. It offers valuable insights into the importance of each feature for the three models we developed: Decision Tree, Random Forest, KNN. By understanding the significance of these features, healthcare professionals can effectively prioritize which values to examine and focus on when assessing patients for CKD. For example, some features like ‘rc’ (red blood cell count), ‘hemo’ (hemoglobin), ‘sc’ (serum creatinine) maintain significant importance across most or all 3 models, suggesting that these values have a much greater impact on CD prediction. Meanwhile, features like ‘PCC’ (pus cell clumps) have very little importance and, as such, may be individually less important in predicting CKD.

The important thing to note here is that this visual provides a quick and easy way to determine some of the most important labs to prioritize in CKD screening as abnormal values in highly important features could serve as early warning signs and lead to sooner detection and intervention. Additionally, while some features show varying levels of importance, suggesting they may not be primary drivers of CKD, they still provide important context in the comprehensive evaluation of the patient. It may be the case that, individually, these features hold less importance, but the model may be capturing complex relationships and interactions with other features that improve its predictive power. Furthermore, as shown in the feature importance chart, different models place significantly different importance on individual features,

further highlighting that just because one model assigns low importance to a feature doesn't mean it isn't significant.

Application of Insights

The insights derived from our analysis are poised to transform CKD screening and management:

Enhanced Screening Programs: Our findings facilitate the refinement of screening programs which allow for the prioritization of resources towards individuals most at risk based on identified key indicators.

Patient Education and Management: By providing targeted education about the importance of managing risk factors such as hypertension and diabetes, patients can take proactive steps to manage their health, potentially staving off CKD.

Conclusion

Our extensive data analysis underscores the potential of data-driven approaches in predicting and managing CKD. By focusing on the early detection of key health indicators, healthcare providers can dramatically improve patient outcomes. This report illustrates how targeted data analysis can inform more effective healthcare strategies, ultimately reducing the prevalence of CKD and enhancing the quality of patient care.