Exploring Predictive Factors for Chronic Kidney Disease

Data Science II

COSC 4337

Submitted to

Dr. Ricardo Vilalta

Submitted by

Joseph Irving (1766731)

Daniel Emami (1390157)

Ryan Nguyen (1897135)

Data Description

Dataset: https://www.kaggle.com/datasets/mansoordaku/ckdisease

The dataset we've used for this project is a collection of features that we intend to use for the prediction of chronic kidney disease (CKD) and was collected over approximately two months from a hospital setting. This multivariate dataset was primarily meant for classification tasks, as indicated by the classification variable, focusing on distinguishing between individuals with and without chronic kidney disease. It includes a total of 400 records with 25 features each. These features include both numerical and nominal types, with 11 numerical attributes describing quantitative measures like age, blood pressure, and various blood tests. There are also 14 nominal values for qualitative properties like the presence of hypertension, diabetes, etc. Our goal is to use these features, along with the class label (CKD for chronic kidney disease and notckd for non-chronic kidney disease) to accurately classify these patients. The raw dataset contains missing values which we intend to correct through the data cleaning process before analysis. This combination of real-world medical data and additional challenges such as missing values makes this dataset an ideal scenario for applying machine learning techniques.

Feature Descriptions

Feature	Description	Datatype
age	Age in years	Numerical
bp	Blood pressure in mm/Hg	Numerical
sg	Specific gravity	Nominal
al	Albumin	nominal
su	Sugar	Nominal
rbc	Red Blood Cells, normal/abnormal	Nominal
pc	Pus Cell, normal/abnormal	Nominal
pcc	Pus Cell clumps, present/notpresent	Nominal

ba	Bacteria, present/notpresent	Nominal
bgr	Blood Glucose Random in mgs/dl	Numerical
bu	Blood Urea in mgs/dl	Numerical
sc	Serum Creatinine in mgs/dl	Numerical
sod	Sodium in mEq/L	Numerical
pot	Potassium in mEq/L	Numerical
hemo	Hemoglobin in gms	Numerical
pcv	Packed Cell Volume	Numerical
wc	White Blood Cell Count	Numerical
rc	Red Blood Cell Count in millions/cmm	Numerical
htn	Hypertension, yes/no	Nominal
dm	Diabetes Mellitus, yes/no	Nominal
cad	Coronary Artery Disease, yes/no	Nominal
appet	Appetite, good/poor	Nominal
pe	Pedal Edema, yes/no	Nominal
ane	Anemia, yes/no	Nominal
classification	Class variable indicating presence of CKD, ckd/notckd	Nominal

Data-Preprocessing

The data preprocessing phase for the kidney disease dataset is fundamental, ensuring the data is structured and ready for analysis. This section outlines the critical steps taken to refine the dataset including cleaning, handling missing values, and reducing dimensionality, which sets a solid foundation for the subsequent analytical tasks.

The first step in our preprocessing was to load the dataset into a pandas DataFrame. We dropped the 'id' column as it does not contribute to our predictive model. Next, we handled any missing values. For numerical columns, we converted non-numeric entries to NaN (Not a Number), and then replaced these NaN values with the mean value of the respective column. This was done for the *pcv*, *wc*, *rc* columns, as well as several additional numeric columns. For categorical columns, we replaced any NaN values with the most frequent value (mode) in the respective column. After handling missing values, we performed a check to ensure that there were no null values left in the DataFrame.

After addressing the issue of outliers and encoding non-numeric columns, we performed feature selection using Recursive Feature Elimination (RFE). This process involves fitting a model to the data, ranking the features, and pruning the least important features until the specified number of features is left. After selecting the most relevant features, we created a new DataFrame with only these selected features and added the target variable *classification* back in. We then displayed summary statistics of the DataFrame using the 'describe' function. Finally, we saved the cleaned and preprocessed dataset to a new CSV file. This dataset was now ready for further analysis or to be used as input for a machine learning model.

Our preprocessing steps ensure that the dataset is clean, free of null values, and only contains relevant features, making it ready for the next steps in our analysis. This thorough preprocessing is crucial to ensure the quality and relevance of the data being used. This combination of real-world medical data and additional challenges such as missing values makes this dataset an ideal scenario for applying machine learning techniques.

Handling Outliers

We opted for using a winsorization method to handle outliers in the dataset. Winsorization is a technique that limits the extreme values in the data to reduce the impact of possible erroneous outliers. Specifically, for our data, we opted to adjust values below the 5^{th} percentile to the 5^{th} percentile and values above the 95^{th} percentile to the 95^{th} percentile for each numeric column.

Considering our options for handling outliers, we considered outright removing the outlier rows or possibly reassigning their values to the mean. Ultimately, we chose the winsorization method for several reasons. One of the most critical reasons for this decision is that this data is medical in nature. In medicine, extreme values can often be a sign of significant clinical findings rather than anomalies. For example, individuals with test results that are extremely high or low might be experiencing severe health problems or advanced stages of a disease, which would be vital for a predictive model with the purpose of diagnosing conditions like chronic kidney disease. By using this method to handle outliers, the dataset retains some of the impact of extreme measurements but limits their influence thus preserving the comprehensiveness of the medical data.

The removal of any records with outliers could lead to the loss of important information, especially since there's a high likelihood that patients with more significant disease states could be represented by those outliers. Furthermore, reassigning those to the mean could potentially mask critical patterns disease progression or severity.

Thus, winsorization appeared to be the perfect balance for reducing the impact of such outliers without fully disregarding the clinical significance of such extreme measurements. This will allow our predictive model to remain sensitive to the nuances of medical data.

Feature Selection

Feature Selection is an important step in the preprocessing phase of any machine learning project that involves high dimensional data. Its purpose is to identify the most relevant features for selection to use in building the model. This results in improved model performance and interpretability. For medical datasets like ours where the accurate prediction of disease presence is critical, this can have a significant impact on the care of the patient and treatment outcomes.

Given that we were dealing with data that had a class label, we considered two main supervised approaches for feature selection. These included filter methods and wrapper methods.

Filter methods operate by evaluating the importance of a feature based on statistical measure such as correlation with the class variable. Features that then meet a certain threshold of significance are selected. This method is more computationally efficient, and they help reduce the risk of overfitting. The downside of filter methods is that they may overlook interactions between other features that may be important for your model.

Alternatively, wrapper methods evaluate the performance of a specific model using a subset of features. In our case, the recursive feature elimination (RFE) works by building models and eliminating features at each step. The recursion stops when a certain number of features have been selected. Unlike the filter method, wrappers can find important interactions between other features selecting those most relevant to the performance of the model. The downside of this method is that they are much more computationally expensive since you are making new models every step of the way.

In our project, we opted to use RFE with a Logistic Regression base model to select the most important features for predicting chronic kidney disease. We believe this method would be the best for reducing dimensionality and focusing on specifically the most significant features that contribute to the target outcome.

Despite the additional computation cost of using RFE, we believe it was the best option for several reasons. For one, the improved model performance and accuracy is a clear benefit. Also, given our small dataset size of only 400 rows, the computation cost of using this method was more manageable. Additionally, in considering the trade-off between speed and accuracy, the critical nature of this task made prioritizing accuracy even more justifiable.

Ultimately, our decision to use the RFE wrapper method for feature selection was driven by its ability to optimize model performance by selecting for the ideal feature subset. The speed increase of using the filter method for our small dataset would have been negligible and not worth the trade-off of more accurate and relevant selected features. This method ensures that the model is trained on the most informative features leading to improved accuracy and better identification of patients with CKD.

It was necessary to use the saga solver in our logistic regression model with RFE to address the challenge of convergence issues. There are several options for solvers that vary on their impact of performance, but we opted for saga as it is useful for high-dimensional data sets since it's designed to converge faster. Additionally, the decision to run 7000 iterations was made to allow for enough time to minimize the loss function and increase the likelihood of convergence.

Features Selected

Correlation between each feature and the target:

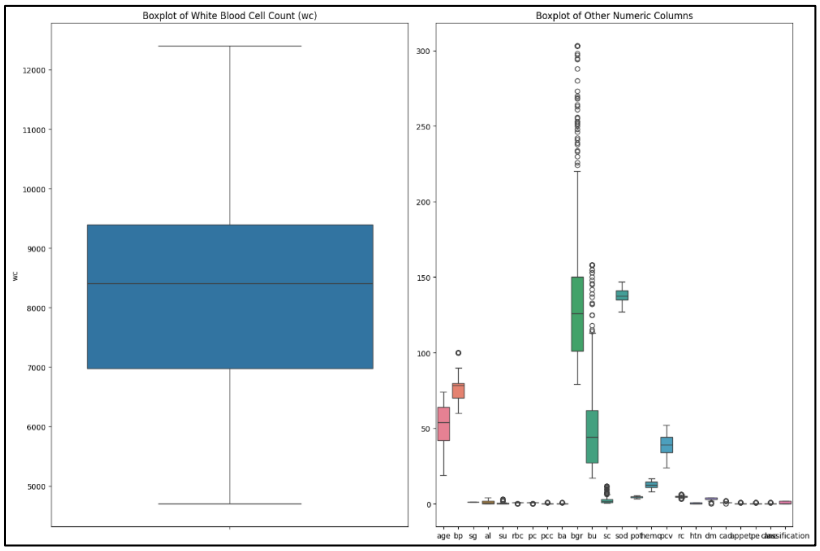
hemo	-0.753794
rc	-0.624580
pc	-0.375154
rbc	-0.282642
pcc	0.265313
su	0.346667
pe	0.375154
appet	0.393341
dm	0.401490
sc	0.455850
htn	0.590438
al	0.600965
classification	1.000000

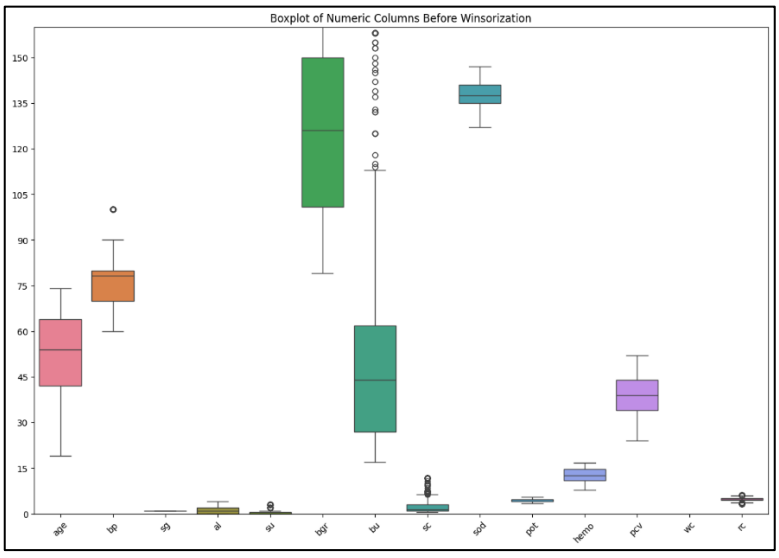
The method output produced 12 features with varying levels of correlation with the target. Among the selected features with strong negative correlations were hemoglobin levels at -0.754 and red blood cell count at -0.625. This aligns with the medical understanding that lower values of these measures are associated with an increased likelihood of CKD. Conversely, strong positive correlations existed with hypertension at 0.590 and albumin at 0.601. Also positively correlated was serum creatinine at 0.457. These findings are intuitive as hypertension can be

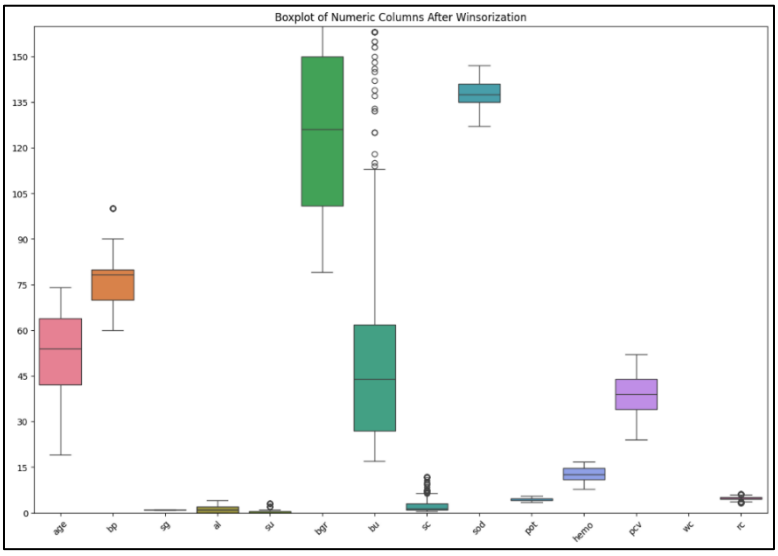
both a cause and consequence of kidney damage, and elevated albumin and serum creatinine are both signs that the kidneys are not functioning properly.

The remaining features selected show a less robust correlation between the feature and the class label. This is likely due to a combination of the RFE method capturing non-linear relationships, feature interactions, and possibly redundancy reduction. Especially in healthcare datasets, it's not uncommon for feature interactions to have a significant impact on the outcome. Individually weakly correlated features when combined with other features may provide important information that contributes to predictive accuracy which RFE evaluates for.

Visualization and Interpretation:







The visualizations highlight *bu* (blood urea) and *sc* (serum creatinine) as key variables, notable for their significant outlier presence both prior to and following the application of Winsorization. This observation led us to employ Winsorization on the dataset, aiming to reduce the impact of these extreme values on our analysis. Outliers, in the context, could possibly signify abnormal levels of blood urea and serum creatinine, which are crucial indicators of kidney function. The elevated levels observed may uncover previously undetected signs of renal dysfunction or impairment, which is particularly relevant to our study's focus on Chronic Kidney Disease (CKD). Conversely, those small indicators of these markers may also indicate underlying health condition or anomalies, which can be backed up by more observation and deeper look into the visuals. We also see that *bgr*, blood glucose random, is not only a big distribution compared to the rest but it also is very high in the numbers which indicates that it has a great impact on our kidney function, so we will need to take a look at patients who have a higher blood glucose random and use that to evaluate whether they have a higher risk of kidney disease or not. The high outliers, above 95^{th} percentile, apply to individuals with abnormally high levels of blood urea and serum creatinine, and low outliers for below 5^{th} percentile that's unusually low.

The decision based on these observations leads us to include or exclude outliers in predictive model requires careful consideration. The outliers may contain valuable information, but they also introduce noise and bias to the model. Using Winsorization methods, we made a great goal to create balance between preserving clinically relevant data and mitigating impact of extreme values. Our model is sensitive in the end to all of these factors, which we have taken account during our initial and final analysis of data, and our three main variables mentioned for visualization play a big role in that. Therefore, we need to make it priority to address these matters when we provide our explanation of these things in our data.