Exploratory Data Analysis on Chronic Kidney Disease

Neeraj Sudheer
PES2UG20CS221
Dept of CSE
PES University

Pranamya P Bhat
PES2UG20CS246
Dept of CSE
PES University

Pratham Manja
PES2UG20CS253
Dept of CSE
PES University

Github repository link:
https://github.com/neerajsudheer/DA-Project_ChronicKidneyDisease

***Abstract*— Over 10% of the world's population is affected with Chronic Kidney Disease (CKD) and millions die every year of kidney failure. In this project, we try to analyze the CKD dataset and come up with effective conclusions. The visualization and analysis of the dataset will help us draw better conclusions on the symptoms as well as the effect of different aspects on the disease.**

Keywords—Chronic kidney disease, Exploratory data analysis.

I. Introduction

The global estimate of chronic kidney disease is around 13.4%. Each year around a million people die from undertreated kidney failure. The main cause of this is the inability to detect the disease at the very early stages. The symptoms and the blood composition levels help us detect CKD at early stages. But most importantly we need to analyze how the disease and the attributes are correlated.We intend to find out the correlation between these health attributes and chronic kidney disease.This way we can allow early detections that facilitate medical interventions immediately. Identify key precursors to chronic kidney disease that can be used for machine learning at a later stage.

The project involves understanding the attributes and the influence it has on the CKD test. Inclusion of graphical representation and other processing for better understanding of the analysis made.

Through this analysis, we intend to understand the amount of dependency each of the attributes has on the disease. This analysis helps us find what the significant cause of the CKD could be. Each of the given attributes in the dataset contribute equally to the causing of the disease. In order to find which factor contributes the most to the disease, what might have caused CKD, etc, we can make use of this analysis. It also talks about how different health issues like anemia, diabetes mellitus, coronary artery disease, pedal edema and hypertension affect the variation in probability of having CKD.

II. Literature survey

One of the existing solutions [1] takes only the correlation between the attributes into consideration. It involves computing the correlation between each of the attributes (or a set of attributes) and the resultant variable. Considering just the correlation as a factor cannot be concluded as the best solution.

The research performed by Wan Chaun Tsai and team [2], tells us about the different aspects that can cause CKD. It is a medicine-based paper where the analysis is solely based on the clinical inferences. The paper provides immense knowledge regarding the medical conclusions that must be taken. It also talks about the inter dependencies of different attributes and the influence it has on the resultant output/predictions. However, it does not include any statistical analysis for the data.

In reference [3], Gulvahid Shaikh and his team have done descriptive statistics of mean and standard deviation (SD). These parameters are used to describe the continuous variables. Frequency and percentages were used to describe the categorical variables. Analysis of variance (ANOVA) was performed adjusting for BUN and glucose. Post-hoc Tukey comparisons compared the means of the CKD groups. Pre and post HD osmol gap comparison was done using the t test. The p values were two sided and were reported to be statistically significant. This reference is the closest to our work. But it considers only the influence of osmol gap on chronic kidney disease. Post researching on CKD, we found that the osmol gap is just one of the factors that affect CKD. We, in our analysis, try to cover every aspect that causes CKD and not just one of them.

Reference [4] tries to analyze the dependence of urinary osmolality and renal outcome on chronic kidney disease. To compare the baseline characteristics according to urine osmolality tertiles, ANOVA with Bonferroni or Kruskal-Wallistest was used for continuous variables, and a

χ2 test was used for categorical variables. Association of the variables were evaluated using univariate and multivariate linear regression analysis. To test the independent variation of urine osmolality on the primary outcome, multivariable cause-specific hazard models were constructed including the significant variables ($p < 0.05$) in univariate analysis, and incremental adjustment was performed. The difference between this analysis and our problem statement is that, we try covering all the basic variables like diabetes mellitus,coronary artery disease, sugar levels, blood pressure, and every other possible symptom/value. Hence we expect to infer a better and symptom specific conclusion than this existing solution.

III. Implementation

A. Preprocessing

We noticed that the dataset required preprocessing as it included several NaN values, null values and outliers.

While preprocessing we have found that there are a significant number of outliers in 'wbc count', 'rbc count', 'pus cells count' which can be inferred by the following table:

Id	0
Age	9
Blood pressure	12
Specific gravity	47
Albumin	46
Sugar	49
Red blood cells	152
Pus cell	65
Pus cell clumps	4
Bacteria	4
Blood glucose random	44
Blood urea	19
Serum creatinine	17
Sodium	87
Potassium	88
Haemoglobin	52
Packed cell volume	70
White blood cell count	105
Red blood cell count	130
Hypertension	2
Diabetes mellitus	2
Coronary artery disease	2
Appetite	1
Pedal edema	1
Anemia	1

We initially tried to drop all the rows with outliers but it resulted in removing around 70% of the tuples. Hence dropping of tuples was not the appropriate way of addressing these attributes with outliers. Hence we have followed the method of imputation. We have replaced these values with the mode of the column values which improved correlation among the attributes.

For the rest of the attributes, when the outliers were analyzed, we found that there isn't a significant number of outliers. Therefore we decided to keep all the outliers.

As far as redundancy is concerned, no duplicates have been found in the dataset.Therefore from the redundancy aspect, the dataset did not require any preprocessing.

B. Libraries used

We use the following libraries for implementation:

- numpy
- pandas
- matplotlib
- seaborn
- scipy.stats

In addition to these main libraries, few might be added as and when needed.

C. Figures and Tables

After a significant amount of preprocessing, we further calculated the correlation between attributes. While plotting the correlation between the attributes we found the following plot:

It can be inferred that blood urea and serum creatinine has a significant correlation along with albumin and serum creatinine.

Acknowledgment

We would like to thank Dr. Prajwala T R for all the support and guidance during this literature survey and analysis regarding the EDA of chronic data analysis.

References

[1] https://www.kaggle.com/code/kianwee/exploratory-data-analysis-chronic-kidney-disease/notebook

[2] Risk Factors for Development and Progression of Chronic Kidney Disease - PMC (nih.gov)

[3] Shaikh G, Sehgal R, Sandhu S, Vaddineni S, Fogel J, Rubinstein S. Changes in osmol gap in chronic kidney disease: an exploratory study. Ren Fail. 2014 Mar;36(2):198-201. doi: 10.3109/0886022X.2013.838052. Epub 2013 Oct 11. PMID: 24111718.

[4] Lee M, J, Chang T, I, Lee J, Kim Y, H, Oh K, -H, Lee S, W, Kim S, W, Park J, T, Yoo T, -H, Kang S, -W, Choi K, H, Ahn C, Han S, H: Urine Osmolality and Renal Outcome in Patients with Chronic Kidney Disease: Results from the KNOW-CKD. Kidney Blood Press Res 2019;44:1089-1100. doi: 10.1159/000502291.