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

The dataset that we chose had following 25 attributes out of which 24 are features and Chronic Kidney Desease is the target attribute

#	Column	Non-Null Count	Type
0	Age (yrs)	391 non-null	float64
1	Blood Pressure (mm/Hg)	388 non-null	float64
2	Specific Gravity	353 non-null	float64
3	Albumin	354 non-null	float64
4	Sugar	351 non-null	float64
5	Red Blood Cells	248 non-null	object
6	Pus Cells	335 non-null	object
7	Pus Cell Clumps	396 non-null	object
8	Bacteria	396 non-null	object
9	Blood Glucose Random (mgs/dL)	356 non-null	float64
10	Blood Urea (mgs/dL)	381 non-null	float64
11	Serum Creatinine (mgs/dL)	383 non-null	float64
12	Sodium (mEq/L)	313 non-null	float64
13	Potassium (mEq/L)	312 non-null	float64
14	Hemoglobin (gms)	348 non-null	float64
15	Packed Cell Volume	329 non-null	float64
16	White Blood Cells (cells/cmm)	294 non-null	float64
17	Red Blood Cells (millions/cmm)	269 non-null	float64
18	Hypertension	398 non-null	object
19	Diabetes Mellitus	398 non-null	object
20	Coronary Artery Disease	398 non-null	object
21	Appetite	399 non-null	object
22	Pedal Edema	399 non-null	object
23	Anemia	399 non-null	object
24	Chronic Kidney Disease	400 non-null	object
25	Blood_Type	400 non-null	object

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 ‘white blood cells’, ’red blood cells’, which can be inferred by the above table

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 mean,median of the column which had continuous values and with mode for the columns with categorical 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

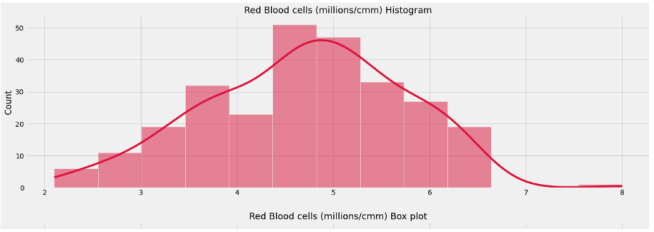
We use the following libraries for implementation:

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

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

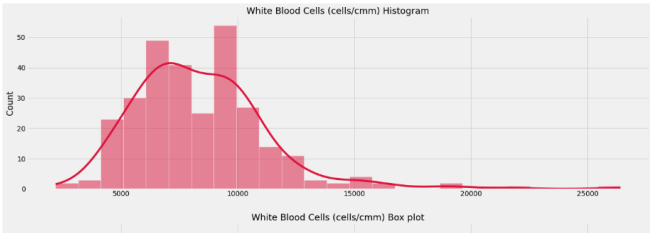
C. Figures and Tables

After a significant amount of preprocessing, we further calculated the correlation between attributes. we further went on doing univariate,bivariate and multivariate



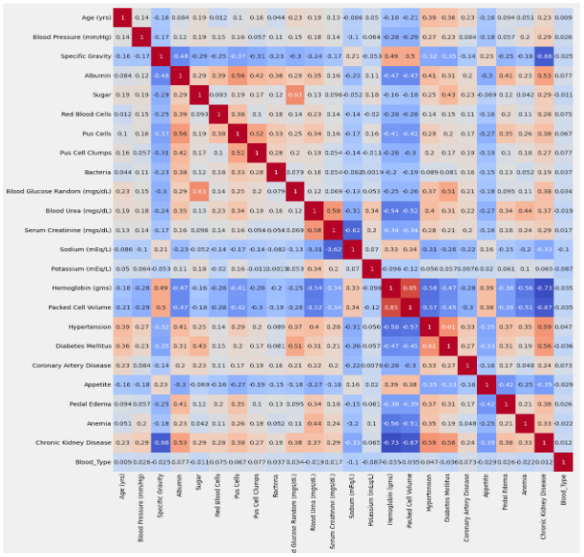
analysis.The figures of that are as follows

from the above plot of univariate analysis of red blood cells attribute we can infer that the count follows nor mal distribution therefore as there are no outliers we can impute the null values with the mean



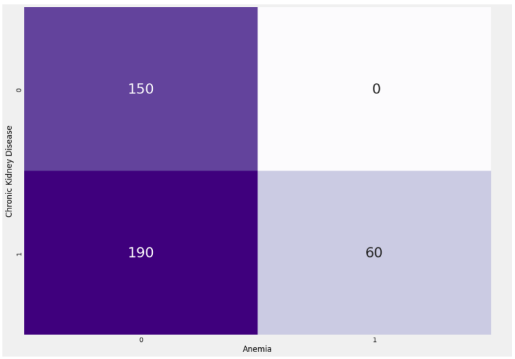
In the same way from the above plot of univariate analysis of white blood cells attribute we can infer that the count does not follow normal distribution and is skewed therefore as there are outliers in this distribution we cannot impute the null values with the mean, therefore we go with median

further when we went on with multi variate analysis we inferred the following correlation plot

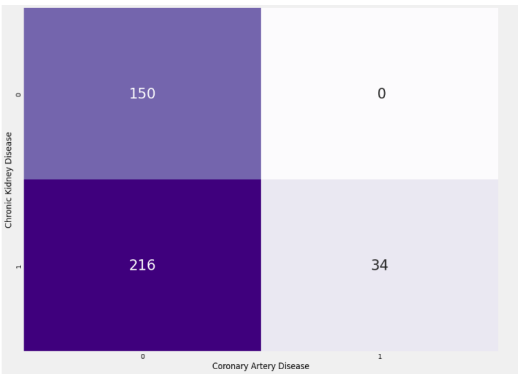


As Hemoglobin and packed cell volume along with Hypertension and Diabetes mellitus have high correlation we can consider them as important features for prediction

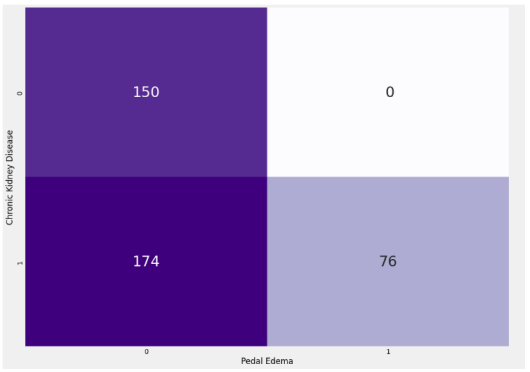
now when we go for bi variate analysis of categorical attributes with the target attribute we inferred the following plots



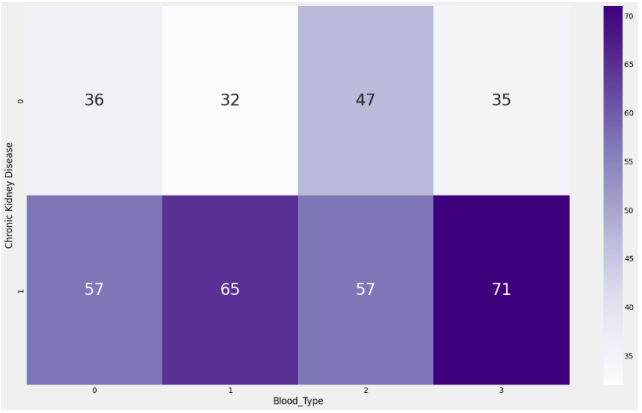
As the Anemia alone cannot distinguish the target attribute we can conclude that the patient having Anemia alone may or may not have CKD



In the same way we can find out that the coronary artery disease cannot classify the target attribute CKD
therefore the person having coronary artery disease may or may not have CKD



In the same way we can find out that the Pedal Edema cannot classify the target attribute CKD
therefore the person not having having Pedal edema may or may not have CKD



Finally when we plot the Blood type with CKD we can infer that more number of patients with blood type O have CKD but it can not be inferred accurately

therefore as we couldn't find out important features directly for our prediction we decided to go on with random forest model to filter out most important features in determination of CKD

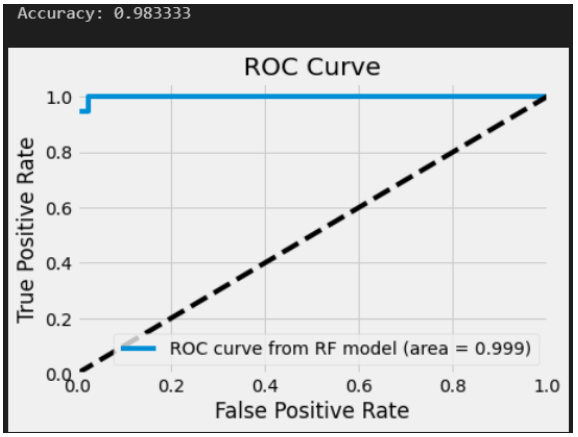
D. Model being used:

After the preprocessing of the dataset, we now proceed to the step of creating a machine learning model to predict kidney disease using the symptoms attribute. Using the SVM algorithm, we train the model to predict if the patient has been affected by the disease or not.

Since it falls under implementation of artificial intelligence in healthcare, utmost accuracy must be aimed. In order to achieve this, a random forest was deployed on this dataset to identify the features that have a higher influence on the

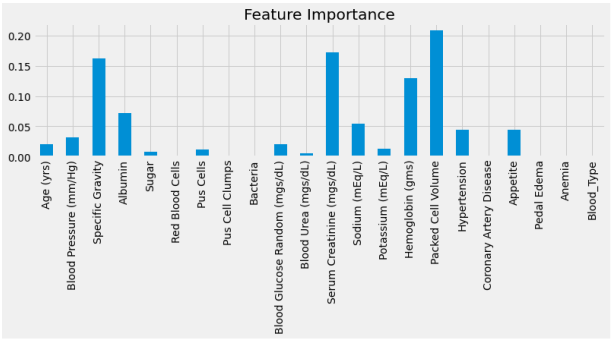
target output. It is important to extract the features that contribute to the prediction of presence of the target variable.

After performing several processing, the results of the random forest model were concluded as the following:



from the above ROC curve it can be inferred that the area under the curve is about 0.99 therefore Random forest is giving good result

Further we used this random forest model to select the important features which contribute significantly in the prediction the result is as follows



By analyzing the above results, we can conclude that featureslike age, sugar, red blood cells, plus cell clumps, bacteria,coronary artery disease, pedal edema, anemia offer a very little impact on the prediction being made. Hence it can be excluded from the analysis and prediction

The processed dataset must now be split in the ratio 70:30 for training and test purposes. Training set is used to help the model in the learning process. The SVM algorithm deployed, assist us train

the machine to classify the data based on the handpicked important attributes of the dataset. Testing is now done using the 30% of data
to find the accuracy and efficiency of the trained model.

After the above computations, it can be concluded that the model has provided us with an accuracy of 97.22% .

IV. CONCLUSION

Data preprocessing plays a major role in any project. It helps in the analysis of the correlation between each attribute and the target and how the changes in the attribute influences the final output obtained.

In the above analysis done on the chronic kidney disease dataset, we have implemented univariate and bivariate analysis to observe the variation of the target value given the symptoms values.

Using random forest method, we conclude on the features that majorly affect the output value. These highly influential attributes are filtered out and the newly formed dataset is used to train the model for prediction purposes. It was noticed that the updated dataset with the important columns,when used to train the model yielded better results with higher accuracy.

As a part of future work, we look forward to implementation of XGBoost for dimensionality reduction and deep learning models for prediction purposes.

V. FUTURE WORK

As seen in the conclusions, the random forest method used as an alternative to dimensionality reduction can be replaced with a regressive boosting method such as XGBoost. This process provides extra mileage and helps the machine learning model to work efficiently. The scope for improvement can further be extended by employing deep learning models of higher processing capacity such as LSTM, or neural networks.

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: 10.1159/000502