Chronic Kidney Disease

December 2020

CS3753

Authors

Ben Baggs

Worked on data analysis, created visuals, worked on modeling and machine learning

--

Tiffany Tran

Worked on data cleaning, created visuals, and included ideas on how to model and do machine learning

--

Next few pages…

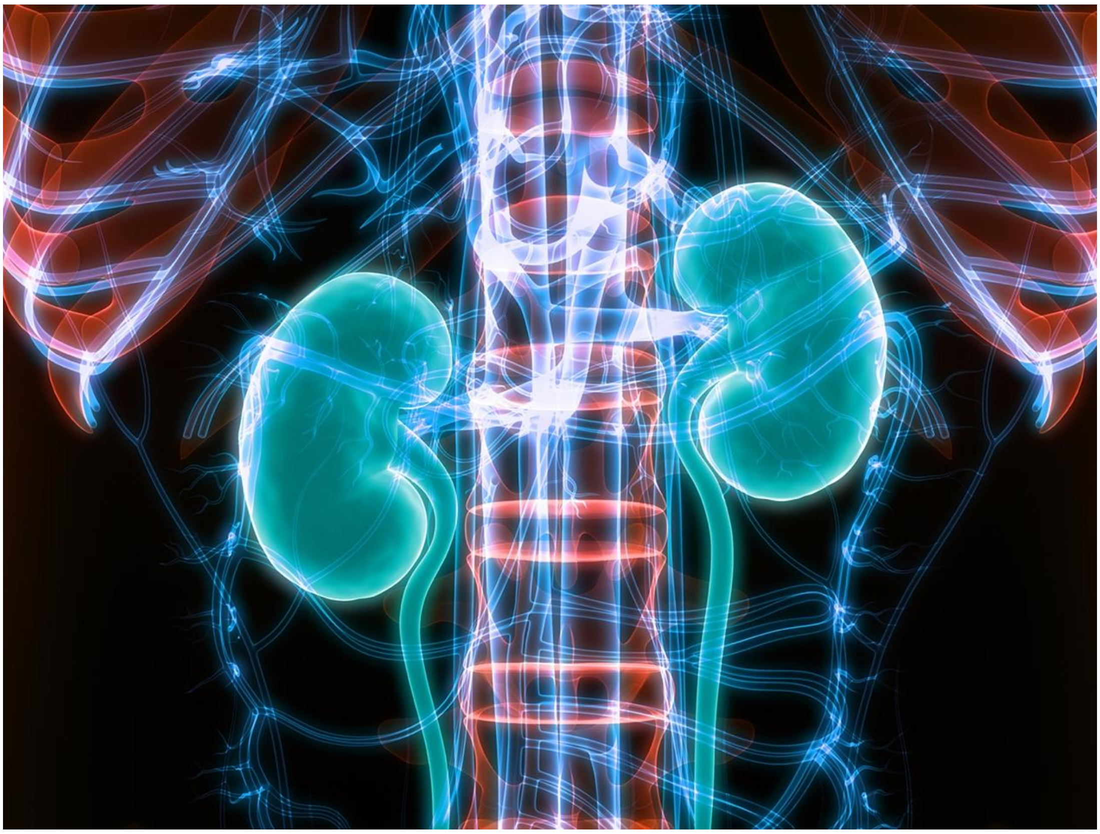
What is Chronic Kidney Disease?

First, let's talk about what your kidneys do normally...

—

Cleaning Data

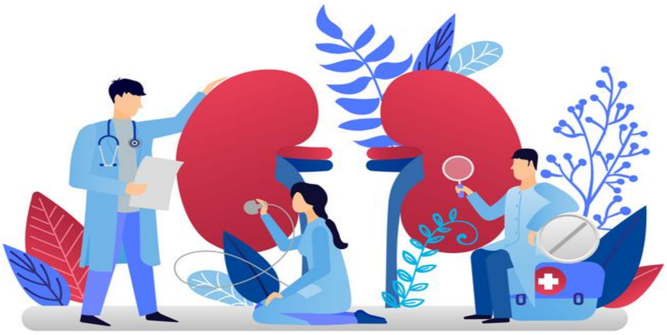
The data that we used has been provided by UCI (6) and Kaggle (5) …



Abstract

Chronic kidney disease (CKD) affects 15% of adults or approximately 37 million people in the United States. Most (9 in 10) adults with CKD do not know they have it. Half of the people with low kidney function who are not on dialysis do not know that they have CKD. (17) The data set we have procured has 25 features which may predict a patient with chronic kidney disease.

The purpose of this project is to utilize machine learning and statistical methodologies to analyze this data set and determine which features are the greatest factors in determining if a patient is likely to have a diagnosis of CKD. Our findings are that some factors overwhelmingly dominate the predictive model in that they are the very definition of the disease, Other features in the model directly correlate to each other, and contribute to the overall probability that the patient has the diagnosis of chronic kidney disease.



What is Chronic Kidney Disease?

First, let’s talk about what your kidneys do normally. Your kidneys are two tiny bean shaped organs each the size of a computer mouse. They are important such as they filter blood, remove waste, and extra fluid from your body. Your kidneys also remove acid that is in your body and they help maintain a healthy balance of water, salts, and other minerals. Healthy kidneys filter about half a cup of blood every minute. (1)

Now, chronic kidney disease (CKD) is a condition where your kidneys cannot filter blood like normal healthy kidneys. Due to this, excess fluid and waste from your blood remain in the body which can lead to many health problems such as heart disease and even induce strokes. (2)

Chronic Kidney Disease has various levels on how serious the disease is. It usually gets worse over time however, treatment for it has shown to slow down its progression. When left untreated, CKD can progress to kidney failure. If kidney failure occurs, then dialysis or a kidney transplant is needed for the patient to survive. (2)

10% of the population globally is affected by chronic kidney disease. (3) In 2017 the global mortality rate increased by 41.5% since the 1990s. (4) As you can see, CKD is very serious, and it is important to for us to know what the indicators are for when you might be susceptible to the disease.

Data Cleaning

The data that we used has been provided by UCI (6) and Kaggle (5). The data has 25 features and 400 rows with each row corresponding to a specific patient.

Without any preprocessing and cleaning of the data. There are 400 rows and 26 columns of data. There are 250 patients known to have been diagnosed with CKD. The other 150 patients are classified as not having the disease. 12 of the features were numeric “float” type data, the remaining 13 were strings, and one column was a unique ID of the patient.

We started with looking at the columns using df.columns. The results of that were a bit hard to understand since the names of the columns were all in abbreviations. (bp, al, su, ba, rbc, etc.)

We started our data cleaning and analysis by changing the column names to be more readable for us.
(sg: Specific_Gravity, pc: Pus_Cell, sod: Sodium, etc.)

Now that the columns were readable, we started looking at what kind of missing values we would have to work with. Overall, we had 1,009 missing values throughout the dataset.
(Can be seen in Figure 1. on next page)

1/3 of the dataset has missing entries. Since our dataset only has 400 rows, removing them would not be a good strategy, therefore we will start by replacing the missing values with their column mean.

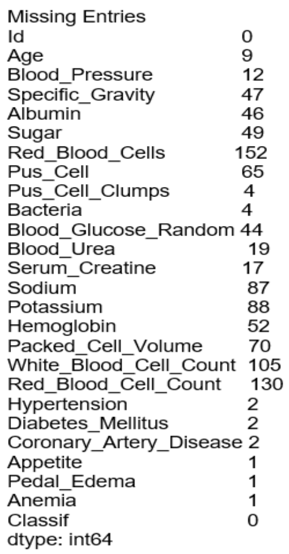
Data Cleaning

Out of 26 columns only 11 columns have the data type of float64. We can start removing the missing entries with these columns since all their values are numbers. Going through each column we took the mean of the column and where there were missing values, we replace those Nan's with the mean.

When inspecting our data closer we found that three columns, that should have been a float64, had a datatype of object. After looking through what kind of values that column produced, we soon realized everything was a number. Hence, we went ahead and changed those columns datatype to be flaot64 and replaced all the Nan's in those columns with their column mean.

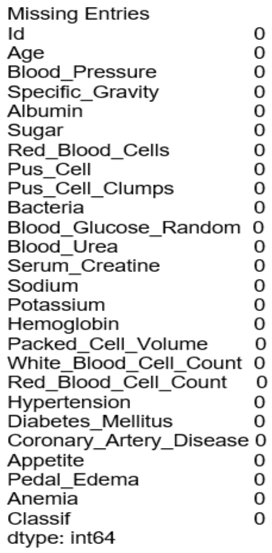
For the rest of the columns that do not have the datatype of float64. We looked through their unique values and found that they all had at most 2-3 distinctive values.

Those values consist of 'yes', 'no', 'unknown', 'abnormal', 'present', etc. We encoded these values and gave them the numbers of 0 or 1. Every value that was positive (yes, present, ckd, normal, good) were given the value of 1. Every value that was the opposite of that was given the value of 0. For those entries that were unknown, we changed those entries to be Nan's instead. Sequentially this allowed us to fill in all the Nan's with the mean of that column.



```
Missing Entries
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_Creatine             17
Sodium                     87
Potassium                  88
Hemoglobin                 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
Classif                     0
dtype: int64
```

Figure 1:Missing Values



```
Missing Entries
Id                         0
Age                        0
Blood_Pressure             0
Specific_Gravity           0
Albumin                    0
Sugar                      0
Red_Blood_Cells            0
Pus_Cell                   0
Pus_Cell_Clumps            0
Bacteria                   0
Blood_Glucose_Random       0
Blood_Urea                 0
Serum_Creatine             0
Sodium                     0
Potassium                  0
Hemoglobin                 0
Packed_Cell_Volume         0
White_Blood_Cell_Count     0
Red_Blood_Cell_Count       0
Hypertension               0
Diabetes_Mellitus          0
Coronary_Artery_Disease    0
Appetite                   0
Pedal_Edema                0
Anemia                     0
Classif                    0
dtype: int64
```

Figure 2:After replacing Missing Values

After doing so, this left us with no missing values in the dataset. *(Shown in Figure 2)*

Now that our dataset has no missing values, we can start by analyzing the data. Plotting each feature against each other would take a while. Thus, we used Seaborn to make a PairGrid to look at all the correlation the columns had with each other. *(Shown in Figure 3.)*

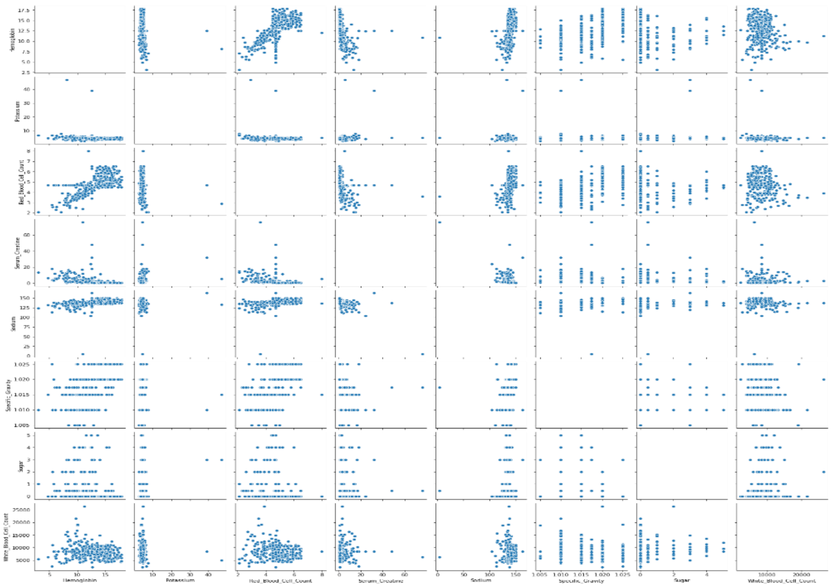


Figure 3: Quick PairGrid graph to see correlation

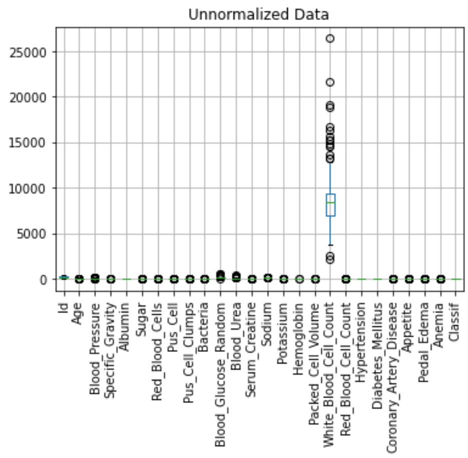


Figure 4: Unnormalized Data

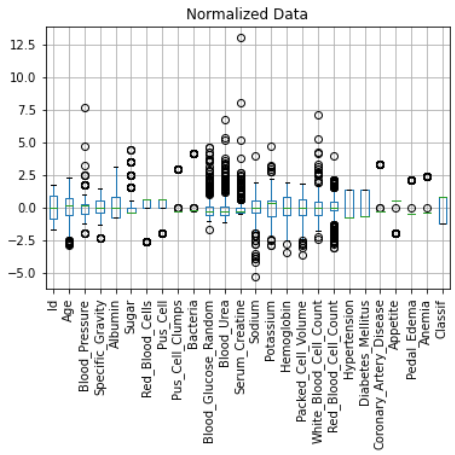


Figure 5: Normalized Data

Data Cleaning

Although the pair grid looks interesting and shows us the correlation between each feature the rest of the features. It doesn't tell us what might correlate with Chronic Kidney Disease.

We then looked at the data and plotted a boxplot to see how everything was in relation to each other. *(Shown in Figure 4.)*

Already you can tell in Figure 4. the boxplot cannot tell us much about the data. We can conclude that there might be a lot of outliers in the column White_Blood_Cell_Count. However, we cannot be sure since our data is not normalized.

To gain more consistency and avoid data anomalies we went ahead and defined a function called zscore that normalized our data. The function follows the normalization formula. In Equation 1. you can see the formula; x is the data point, and we subtract the mean from z and divide by the standard deviation.

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \text{Mean}$$
$$\sigma = \text{Standard Deviation}$$

Equation 1: Normalization Formula

After running our dataset through zscore. We plotted our results with another boxplot. *(Shown in Figure 5.)*

This new boxplot looks much better! Normalizing the data was good since now we can see every feature. Now that we have normalized the data and plotted it. There is something that is prevalent in the graph.

Outliers

Outliers are observations that lies outside the normal range from the sample data. However, when dealing with data, it generally is up to the analyst to decide what is an outlier and what isn't. (7)

We decided to look at the max and min of each column. We took those values that we found and researched them to see if they were valid or not. In most cases what seemed to be outliers based on the definition, of what lies outside the normal range, ended up not being outliers after we had researched them.

Blood_Pressure for example looks like it could have a lot of outliers since the normal range based on the boxplot is near 0.0. While the outliers are further away from the norm. After inspecting it more the maximum value of Blood_Pressure was around 180 mm Hg. The minimum value in the column was 50 mm Hg.

While in the boxplot graph *(Shown in Figure 5.)* Blood_Pressure looks to have a few outliers; this is not the case.

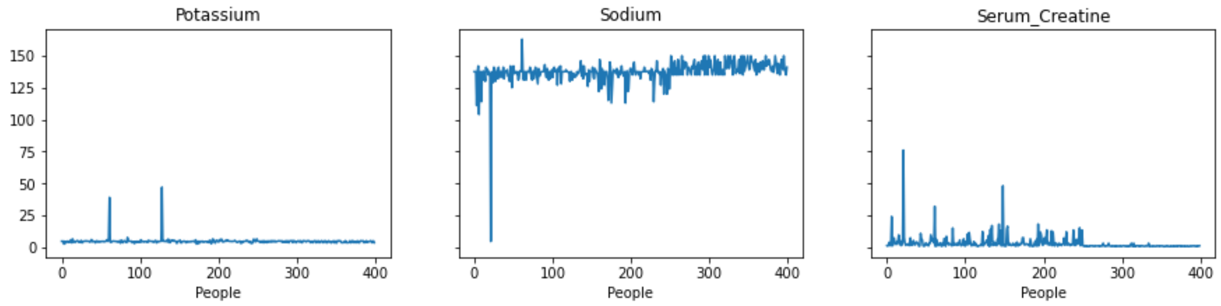


Figure 6: Data with Outliers

Outliers

When we look at our lowest recorded blood pressure in the dataset 50 seems way to low. Especially when it's widely known that the normal range for blood pressure needs to be less than 120/80 mm Hg. (8)

However, the American Heart Association states that those who have a resting heart rate slower than 60 BPM tend to be physically active adults or they are athletes. (9)

Now if we look at our highest blood pressure recorded in the dataset being 180. That seems way too far from 120 mm Hg so it must be an outlier.

Nevertheless 180 mm Hg is a valid blood pressure reading as well. It's not normal but it is possible. A reading higher than 180 mm Hg is considered critical hypertension. (9)

When looking for outliers, we find it very important to keep all data that is realistic in the dataset. What we are looking for are those values that are impossible.

After going through all the features that we have looking for abnormal data. We ended up finding three columns with some unusual values.

Potassium and Sodium had abnormal outliers *(Shown in Figure 6.)* where upon further inspection, we found that these values were not valid.

In the Potassium graph *(Shown in Figure 6.)* there are two spikes we saw. We investigated it and the two spikes had values of 39 and 47 millimoles per liter (mmol/L).

Normal blood potassium levels are in the range of 3.6 to 5.2 millimoles per liter (mmol/L). (10)

Looking at the other values in the column this seems to be true. However, we had to find out more information, so we researched to see what the highest level of potassium recorded.

The highest potassium level that has ever been recorded was in 2005 where the patient had a reading of 14.0 millimoles per liter (mmol/L). (11) This makes it impossible for someone to have a reading of 39 and 47 millimoles per liter (mmol/L).

Sodium had one outlier that we found to be not valid. *(Spike shown in Figure 6.)* The patient had a reading of 4.5 milliequivalents per liter (mEq/L). According to the MayoClinic organization, a normal reading for sodium are levels between 135 and 145 milliequivalents per liter (mEq/L). (12)

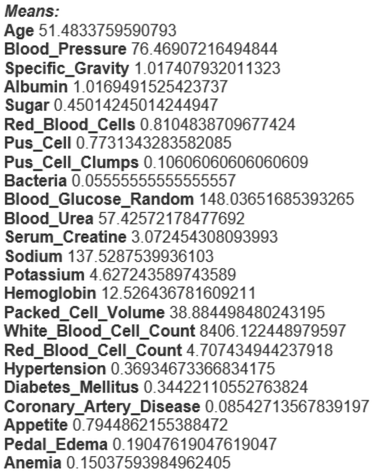
Due to these case findings, we decided perhaps the outliers in Potassium and Sodium were typos in the system. We went ahead and removed those values and replaced them with the mean of their column.

Serum_Creatine was an odd column to work with. After going through the data, we looked at the maximum and minimum values. What we found was shocking and a bit disappointing.

For results of serum creatine, we found that it could be measured in milligrams per deciliter or micromoles per liter. Ranges may be 0.84 to 1.21 Mg/dL (74.3 to 107 mcmol/L). (13)

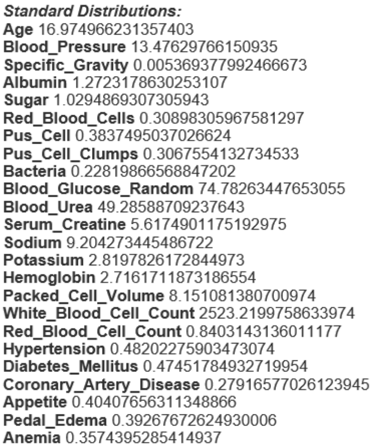
We believe our data set has a mixture of both Mg/dL and mcmol/L. Due to this we decided not to do anything with the column yet. It later will be removed before we begin modeling.

Results and Discussion



```
Means:
Age 51.4833759590793
Blood_Pressure 76.46907216494844
Specific_Gravity 1.017407932011323
Albumin 1.0169491525423737
Sugar 0.45014245014244947
Red_Blood_Cells 0.8104838709677424
Pus_Cell 0.7731343283582085
Pus_Cell_Clumps 0.10606060606060609
Bacteria 0.05555555555555557
Blood_Glucose_Random 148.03651685393265
Blood_Urea 57.42572178477692
Serum_Creatine 3.072454308093993
Sodium 137.5287539936103
Potassium 4.627243589743589
Hemoglobin 12.526436781609211
Packed_Cell_Volume 38.884498480243195
White_Blood_Cell_Count 8406.122448979597
Red_Blood_Cell_Count 4.707434944237918
Hypertension 0.36934673366834175
Diabetes_Mellitus 0.34422110552763824
Coronary_Artery_Disease 0.08542713567839197
Appetite 0.7944862155388472
Pedal_Edema 0.19047619047619047
Anemia 0.15037593984962405
```

Figure 7: Means of the Features



```
Standard Distributions:
Age 16.974966231357403
Blood_Pressure 13.47629766150935
Specific_Gravity 0.005369377992466673
Albumin 1.2723178630253107
Sugar 1.0294869307305943
Red_Blood_Cells 0.30898305967581297
Pus_Cell 0.3837495037026624
Pus_Cell_Clumps 0.3067554132734533
Bacteria 0.22819866568847202
Blood_Glucose_Random 74.78263447653055
Blood_Urea 49.28588709237643
Serum_Creatine 5.6174901175192975
Sodium 9.204273445486722
Potassium 2.8197826172844973
Hemoglobin 2.7161711873186554
Packed_Cell_Volume 8.151081380700974
White_Blood_Cell_Count 2523.2199758633974
Red_Blood_Cell_Count 0.8403143136011177
Hypertension 0.48202275903473074
Diabetes_Mellitus 0.47451784932719954
Coronary_Artery_Disease 0.27916577026123945
Appetite 0.40407656311348866
Pedal_Edema 0.39267672624930006
Anemia 0.3574395285414937
```

Figure 8: STD of the Features

Preliminary analysis of the data indicated that the mean age of our patients was 51.5 years of age and had a standard deviation of 17 years. *(Shown in Figure 8)* The mean diastolic blood pressure was below average for the mean of our age group indicating our data sample was relatively healthy.

The specific gravity of urine was striated into 5 unique values, and the mean *(shown in Figure 7)* of the data set was within normal range. The mean blood urea level was abnormal indicating that the dataset might have more patients with CKD than without.

Blood Glucose levels were just high of average indicating the dataset might have slightly more diabetic or pre-diabetic patients. The mean of serum creatinine levels was abnormally high for this dataset.

Upon further analysis it was discovered the serum creatinine levels were measure by two different standards and it was noted that this should be taken into consideration for all data experiments moving forward.

The sodium and potassium means for the dataset where within normal values. The hemoglobin mean levels skewed slightly towards the low side of normal indicating a concentration of anemic people in the dataset.

The hemoglobin correlated directly with the low mean red blood cell count of our dataset. The features packed cell volume, white blood cell count, sugar, albumin, were all within normal mean values. Approximately 37% of the patients in our data set were diagnosed with hypertension and 34% of our patients were diagnosed diabetics.

Further analysis into the data set revealed that 250 of the patients were diagnosed with CKD as opposed to the 150 patients without. This imbalance of those diagnosed the CKD and those without, explains any of the mean values in the dataset that fell to the lower side or outside of the normal range.

A nested iteration of Pearson correlation coefficients of all float values yielded expected correlations of related values. For example, hemoglobin and red blood cells directly correlated with a linear coefficient of nearly 0.7, serum creatinine and sodium had a coefficient of – 0.6, and specific gravity of urine and hemoglobin in blood has a coefficient of 0.5.

All other results of this experiment were unremarkable. A matrix of scatterplots revealed that some portions of the data was striated into nearest values. For example, there were only 5 unique values for specific gravity of urine, 5 unique values of albumin, and 7 unique values for blood sugar as a percentage. As we only found a few values like this we opted to utilized multiple linear regression in our model however isolating the striated data would be a good candidate for clustering.

With all the data in numerical form and with the intention of negating unknown values by replacing them with a column mean and normalizing them we chose multiple linear regression to analyze our data. The initial results were quite surprising in that one coefficient massively dominated the entire set. Specific gravity of urine had an overwhelming correlation to CKD. *(Shown in Figure 9)*

Specifically, low specific gravity of urine or diluted urine absolutely negated all the other factors in our dataset. However, this metric is the definition of CKD.

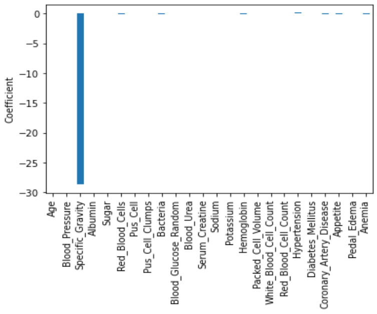


Figure 9: Linear Regression Model w/ Specific_Gravity

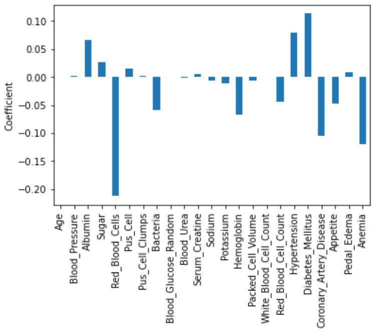


Figure 10: Linear Regression Model w/o Specific_Gravity

Furthermore, diabetes finds itself as a significant factor once again as well. However, with the SHAP model *(shown in Figure 12)* we finally see serum creatinine have an impact when in conjunction with diabetes and blood related counts. It seems that a creatinine level greater than 2.5 in conjunction with diabetes and lack of red blood cells most directly correlated to CKD according to our model.

Results and Discussion

To dig deeper we removed specific gravity as a feature in the linear regression and ran the regression again. *(Shown in Figure 10)* With the specific gravity removed we saw a more balanced equation of factors.

Specifically, we saw that red blood cells count, followed by anemia, and diabetes had the highest correlations with the CKD diagnosis.

We also discovered that age, blood pressure, and surprisingly blood urea did not have significant correlation to CKD in perspective of the other data. Perhaps due to CKD patients receiving dialysis. Also, sodium, potassium, and creatinine were relatively insignificant in comparison to the other data.

For out next analysis we opted for the SHAP library to explain our results. Once again red blood cells and therefore hemoglobin have some of the highest values in the model that shift the result of not having CKD to having CKD.

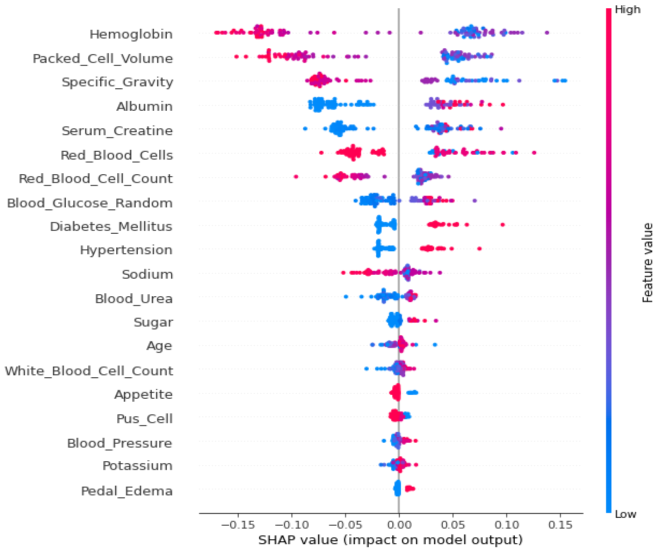


Figure 11: SHAP Values

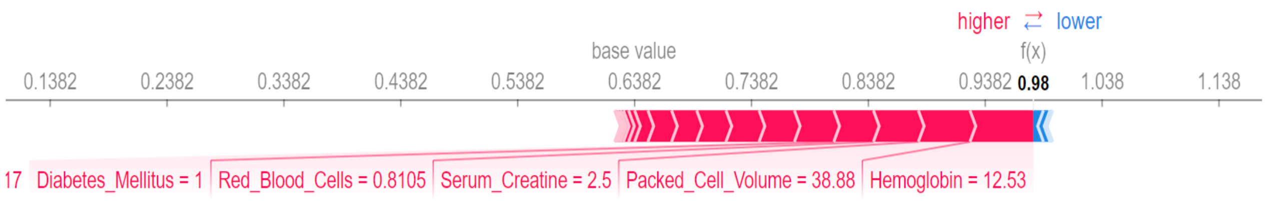


Figure 12: SHAP Model

Conclusion

Our analysis of the data does not separate cause from effect it simply provides the most reliable indicators of a person with CKD.

As we saw in the data analysis diluted urine is overwhelmingly the largest indicator of kidney disease and our research on the internet concurs. It makes sense that a kidney that does not efficiently filter waste products out of the blood will produce diluted urine with an absence of waste product.

According to rnceus diluted urine occurs under 3 circumstances; diabetes causing a decrease in anti-diuretic hormone, damage to the kidney's tubules affecting the kidneys abilities to re-absorb water, and renal failure. It is also no coincidence that according to davita.com diabetes is the number one cause of CKD. (15)

According to the article, "Diabetes is a disease that affects the body's ability to produce or use insulin. When the body turns the food eaten into energy (also called sugar or glucose), insulin is used to move this sugar into the cells. If someone produces little or no insulin, or if the body cannot use the insulin (insulin resistant), the sugar remains in the bloodstream instead of going into the cells. Over time, high levels of sugar in the blood damage tiny blood vessels throughout the body including the filters of the kidneys. As more damage occurs to the kidneys, more fluid and waste remain in the bloodstream instead of being removed.'' (14)

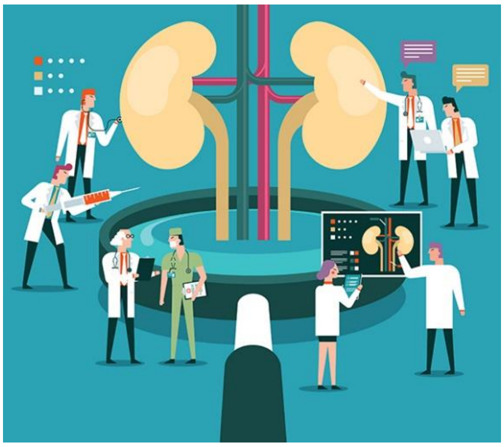
Further research showed that healthy kidneys make a hormone called erythropoietin (EPO). EPO sends a signal to the body to make more red blood cells. (15)

If your kidneys are not working properly and they can't make enough EPO and your body does not make enough red blood cells. This lack of blood cells in turn can cause decreased hemoglobin and this in turn causes anemia. It is no wonder that after urine specific gravity, red blood cell count, hemoglobin, and anemia features were our next highest factors on the dataset.

High blood pressure, coronary artery disease, hypertension, and pedal edema are all interrelated to symptoms. These symptoms were the next highest group of indicators of CKD in our model. (16)

Finally, damaged kidneys may release too much renin, which can lead to high blood pressure. Whether the CKD is causing the symptom, or the symptom is causing the CKD our analysis shows that the higher the presence of these symptoms the greater the chance of having a CKD diagnosis.

Surprisingly, our findings did not show as high of a correlation of waste products in the blood having as big of an impact on the likely diagnosis. Blood levels of potassium, and sodium had a negligible impact to the diagnosis of CKD in our model. We may speculate this is due to dialysis treatment or perhaps they just do not fluctuate outside of normal values enough to distinguish a patient with or without CKD.



Author Contributions

Ben Baggs

Worked on data analysis, created visuals, worked on modeling and machine learning. Also wrote article.

Tiffany Tran

Worked on data cleaning, created visuals, and included ideas how to do modeling and machine learning. Also wrote article.

References

1. https://www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-they-work
2. https://www.cdc.gov/kidneydisease/basics.html
3. https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease
4. https://www.medscape.com/answers/238798-105274/what-is-the-global-prevalence-of-chronic-kidney-disease-ckd
5. https://www.kaggle.com/mansoordaku/ckdisease
6. https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease
7. https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm#:~:text=An%20outlier%20is%20an%20observation,random%20sample%20from%20a%20population.&text=Examination%20of%20the%20data%20for,often%20referred%20to%20as%20outliers.
8. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings
9. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

References

10. https://www.mayoclinic.org/symptoms/low-potassium/basics/definition/sym-20050632#:~:text=Normally%2C%20your%20blood%20potassium%20level,and%20requires%20urgent%20medical%20attention.
11. https://acutecaretesting.org/en/journal-scans/a-record-breaking-serum-potassium-concentration#:~:text=Now%20a%20recently%20published%20case,a%20patient%20who%20has%20survived.
12. https://www.mayoclinic.org/diseases-conditions/hyponatremia/symptoms-causes/syc-20373711#:~:text=A%20normal%20blood%20sodium%20level,Certain%20medications.
13. https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646#:~:text=Results%20of%20the%20creatinine%20blood,and%20women%2C%20and%20by%20age.
14. https://www.rnceus.com/ua/uasg.html#:~:text=In%20these%20diseases%2C%20damage%20to,gravity%20between%201.007%20and%201.010
15. https://www.davita.com/education/kidney-disease/risk-factors/diabetes-is-the-leading-cause-of-chronic-kidney-disease
16. https://www.kidneyfund.org/anemia/
17. https://www.cdc.gov/kidneydisease/publications-resources/2019-national-facts.html#:~:text=With%20chronic%20kidney%20disease%20(CKD,disease%20and%20high%20blood%20pressure.