Constructing a Classification Model to Predict Chronic Kidney Disease in Patients

Hank Magan

10/26/2021

Abstract

Chronic kidney disease (CKD) is a potentially fatal condition characterized by eventaul kidney failure. This report focuses on a variety of factors which may contribute to the disease, such as blood pressure and red blood cell count. The dataset used in this report is tailored around classification, and as such a KNN classification algorithm was applied to ultimately produce a predictive model for the presence of CKD, based on the various factors explained above. The end product of this study is an 83% accurate model for predicting CKD.

Introduction

Chronic kidney disease (CKD), is a type of kidney disease in which there is gradual loss of function in the kidneys. Again, the process is gradual, and can take anywhere from a few months to a few years in which the symptoms progressively worsen. In 2016, a whopping 753 million people globally were affected by the disease, causing 1.2 million deaths just the year prior. Needless to say, the disease is unfortunately fairly common and a leading cause of death in the United States.

The kidneys are responsible for a variety of things, including removing wastes and toxins, regulating blood pressure, producing red blood cells, and more. With that said, they primarily function to remove waste from the blood and to transform it into urine, through which this waste is removed from the body. When the kidneys eventually shut down as in CKD, these functions gradually become less and less effective, which manifests itself in a variety of symptoms. For example: edema, fatigue, vomiting, loss of appetite, anemia, and even confusion. Additionally, due to the build-up of urea, one symptoms is so-called uremic frost, a term used to describe the crystallized urea deposits that accumulate on those with CKD when they sweat. These symptoms are generally seen later in its progression, as there are initially little to no symptoms, making it difficult to detect early on.

The causes of CKD are numerous. The most common causes are diabetes and hypertension, however there are additional causes such as glomerulonephritis, and polycystic kidney disease. The disease is also linked genetically, and therefore family history can be a risk factor. To reduce the risk of CKD, those with the mentioned causes should limit the causes as much as reasonably possible. CKD itself is not necessarily fatal, although it increases the risk of cardiovascular disease, which is the leading cause of death in the US. In fact, the most common cause of death in people from CKD is cardiovascular disease.

The diagnosis of CKD can be done in different ways, however the most common method is to screen those who are at-risk (family history of CKD, have hypertention, etc) via a urine sample. Diagnosis can also be performed using an ultrasound to identify things like decreased kidney size and cortical thinning, two common indicators of CKD. Once diagnosed, initial treatment usually involves medication to reduce factors such as blood pressure, blood sugar, and cholesterol in the bloodstream. Further treatments include altering diet, and taking vitamins. These are generally steps taken to slow the progression of the disease, not necessarily to cure it. Once it progresses to stage 4, it is generally agreed upon that patients should be referred to a nephrologist. By stage 5, kidney replacement therapy is usually required. This can be done through dialysis or transplantation. A transplant increases the likelihood of survival more, however has potential short-term complications due to surgery. Dialysis can be performed at home (home hemodialysis), or through three-times-a-week hemodialysis and peritoneal dialysis.

Methods

The first thing to do is, as always, to perform EDA on the dataset. The data is then cleaned, which includes renaming variables and handling missing data. Next, the data is examined a little closer using graphical representations to better understand how the data relates to itself. Various graphs are plotted. Lastly, a KNN algorithm is run to create a predictive model for whether a patient has CKD or not. The results of this model are then to be displayed using a cross table.

Results

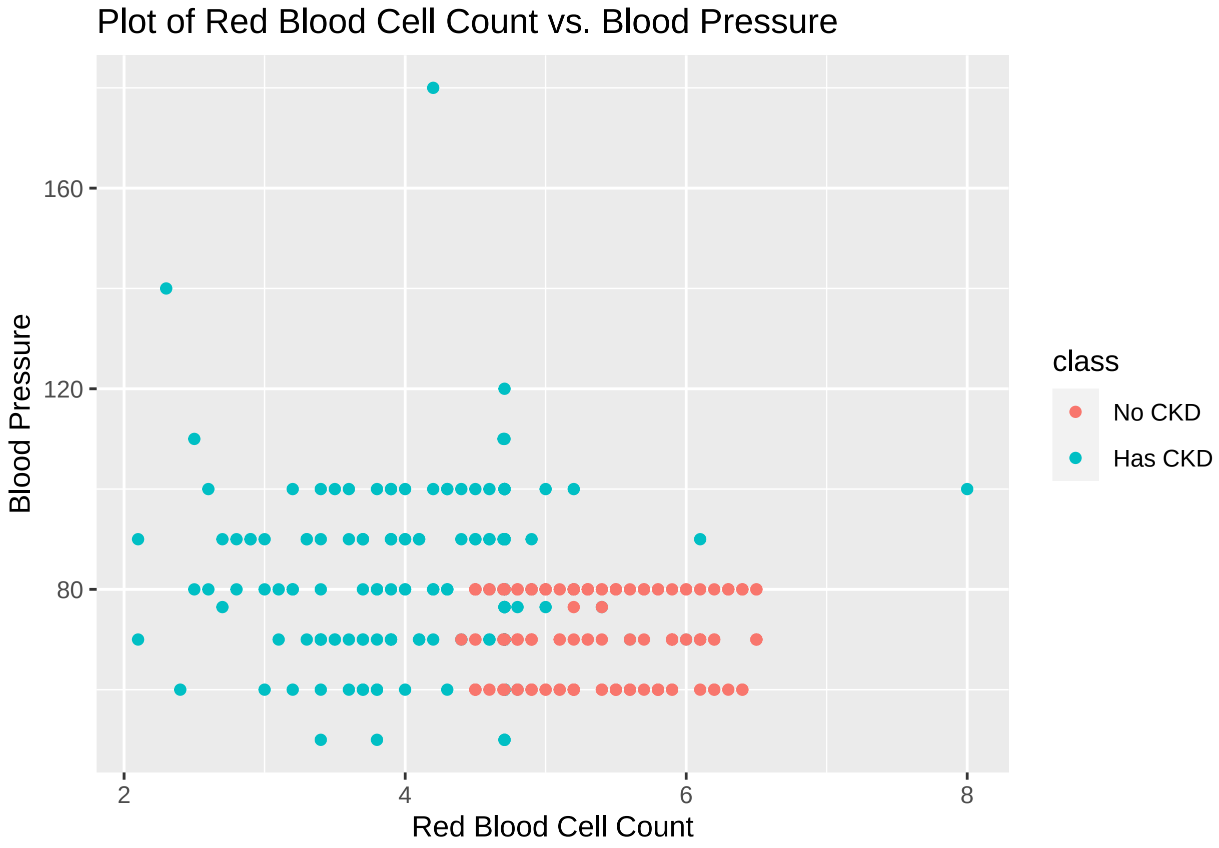
Before conducting any analysis, the data should be briefly examined and cleaned if necessary. The first thing to do is to look at the column names, and make changes if necessary. In the case of this data, all of the variable names made sense and were concise, and therefore unchanged. For reference, the variables and their significance are below.

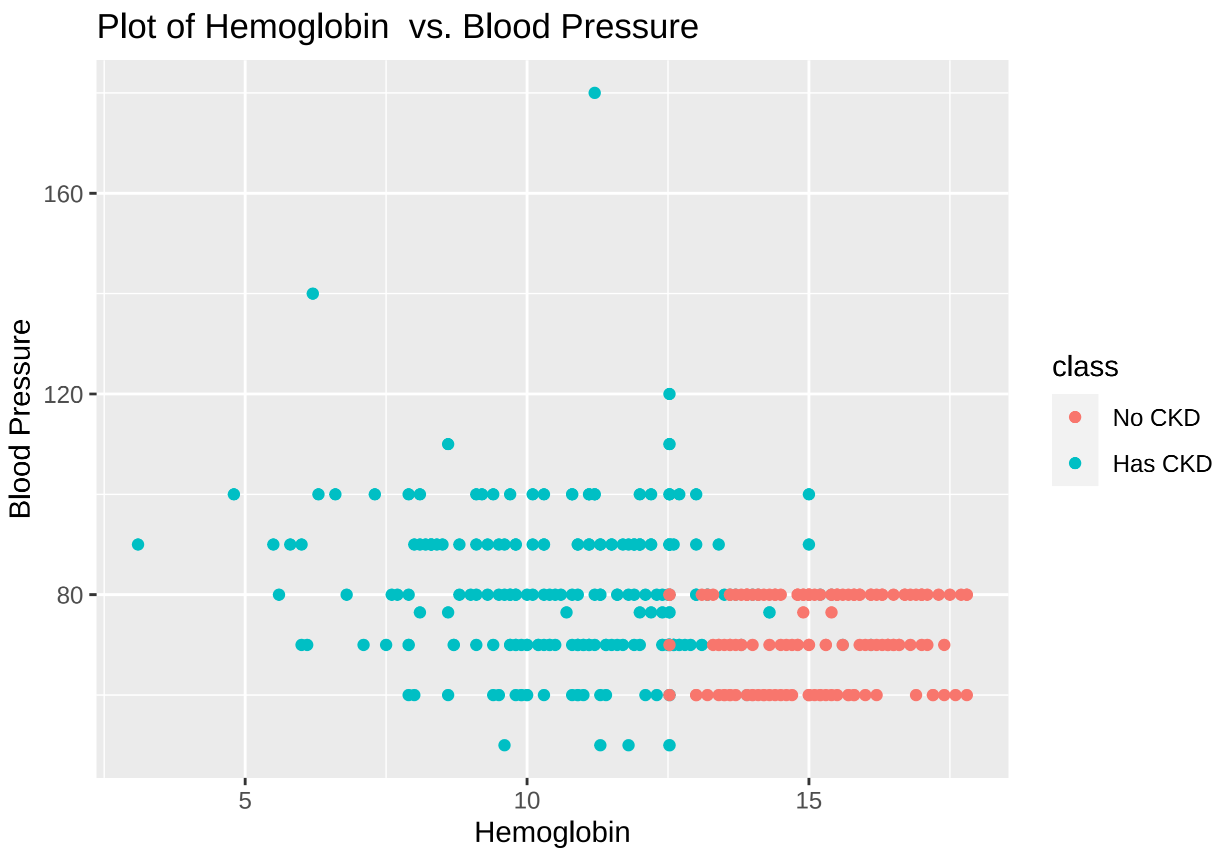
1. age - age
2. bp - blood pressure
3. sg - specific gravity
4. al - albumin
5. su - sugar
6. rbc - red blood cells
7. pc - pus cell
8. pcc - pus cell clumps
9. ba - bacteria
10. bgr - blood glucose random
11. bu - blood urea
12. sc - serum creatinine
13. sod - sodium
14. pot - potassium
15. hemo - hemoglobin
16. pcv - packed cell volume
17. wc - white blood cell count
18. rc - red blood cell count
19. htn - hypertension

20. dm - diabetes mellitus
21. cad - coronary artery disease
22. appet - appetite
23. pe - pedal edema
24. ane - anemia
25. class - class

Using the `glimpse` command gives a nice overview of the data, indicating that there is a total of 400 records. Additionally, running this command indicates two things: there is plenty of missing data that should be resolved, and there is plenty of boolean data that should be converted to 0's and 1's for the sake of the analysis. Taking a closer look using the `summary` command, the amount of missing data per column can be observed. Some columns were hardly missing data, such as age (missing 9 values), while some were missing a significant portion of the data, such as the red blood cell description (missing 152 values), with most falling somewhere in between. For numerical data, missing values were replaced with their respective mean; for boolean data (which was first converted to 0's and 1's), missing values were replaced with 0's.

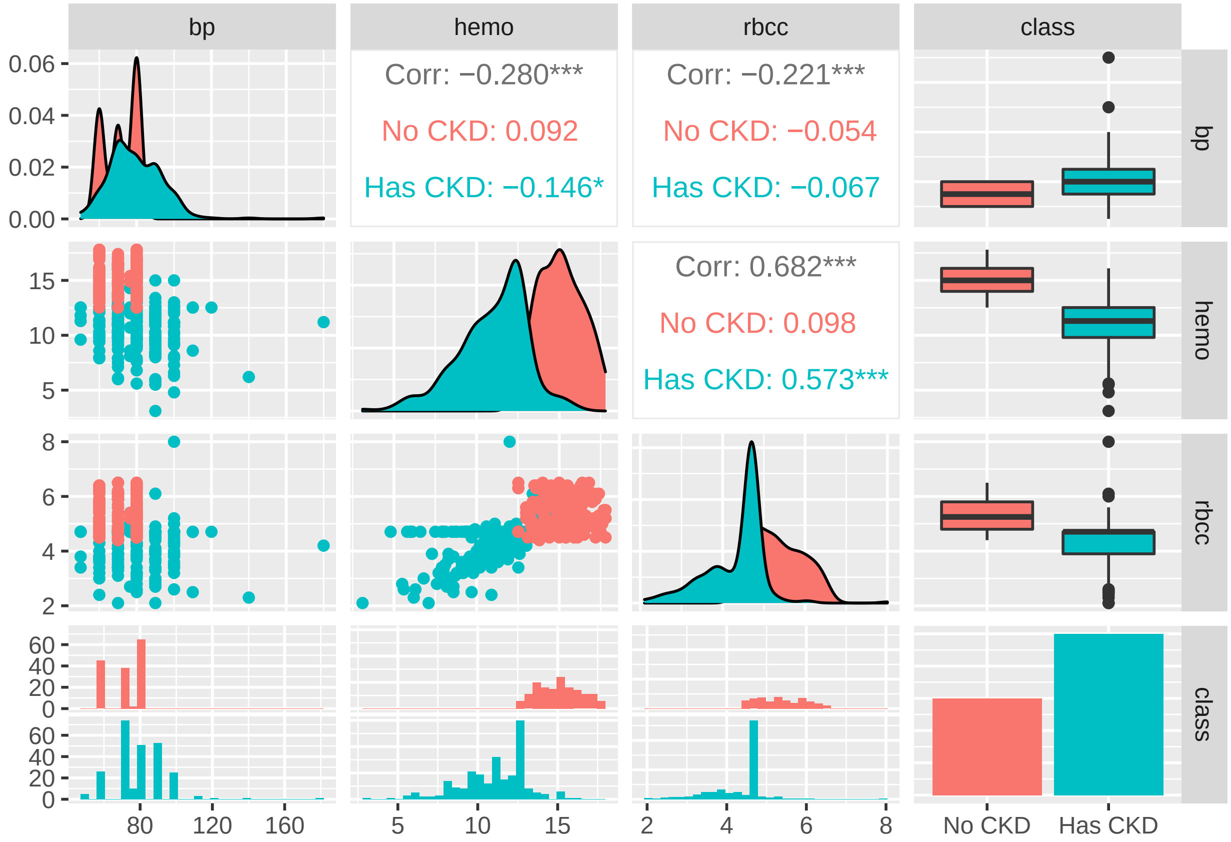
With that, the data is finally clean. Before conducting classification on the data to build a predictive model, some of the data will be examined graphically. As outlined in the introduction, one function of the kidney is to produce red blood cells. Additionally, a symptom of CKD is anemia (lack of hemoglobin). With that said, examine the following graphs:





Based on the plots, it seems that low red blood cell count and low levels of hemoglobin are both associated with CKD. The association is rather well-defined, in fact. This looks promising from a classification standpoint. Additionally, a pairwise table depicting the data can be seen below, which can provide a better understanding of the relationships between variables. It also includes correlation values between variables.

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



With the data properly explored, the actual machine learning aspect of this analysis may be conducted. As discussed previously, this study uses the KNN algorithm for classifying the dataset. This process is done over multiple steps. Firstly, the data is separated into test and training data. For this study, 70% of the data was dedicated to training the model, while 30% was dedicated to testing. Next, a KNN algorithm was run using this training data. When finding k, which specifies the number of neighbors to consider when classifying a data point, various values were used and compared to determine what would provide the best results. Lastly, a so-called cross table was created to demonstrate the model's accuracy in determining CKD diagnosis in the testing data. The results of this analysis are shown below.

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  137
##
##
##                 | data_pred
## data.testLabels |         1 |         2 | Row Total |
## ----------------|-----------|-----------|-----------|
##               1 |        46 |         5 |        51 |
##                 |     0.902 |     0.098 |     0.372 |
##                 |     0.719 |     0.068 |           |
##                 |     0.336 |     0.036 |           |
## ----------------|-----------|-----------|-----------|
##               2 |        18 |        68 |        86 |
##                 |     0.209 |     0.791 |     0.628 |
##                 |     0.281 |     0.932 |           |
##                 |     0.131 |     0.496 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |        64 |        73 |       137 |
##                 |     0.467 |     0.533 |           |
## ----------------|-----------|-----------|-----------|
##
##
```

Conclusion

This table indicates that the model, over a testing set of 137 instances, predicted correctly 83% of the time. This is quite a decent model. The model predicts false negatives about 4% of the time, and false positives around 13% of the time. With that said, whether this model should be trusted as a reliable way to diagnose CKD is questionable; 17% is not a low enough chance for error for the model to be consistently trusted. However, that is not to say that this model is useless. It may still be useful in some applications. Furthermore, this analysis suggests that CKD is fairly predictable in its symptoms. More specifically, hemoglobin and red blood cell count were identified as relatively strong indicators of CKD.

References

1. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease
2. https://en.wikipedia.org/wiki/Chronic_kidney_disease
3. https://www.cdc.gov/kidneydisease/basics.html