Analysis of risk factors for chronic kidney disease

Authors: Shawn Li, Meng Chen
The Johns Hopkins University
Apr 26, 2022

Abstract:

This paper is to analyze what factors affect people's chronic kidney disease. Found risk factors as signals to remind people. We used logistic regression and random forest to do this classification research topic.

1. Introduction

The kidneys are essential to people's lives, they filter all the blood of the body every 30 minutes, and remove wastes, toxins, and excess fluid. They also help control blood pressure, stimulate the production of red blood cells, keep bones healthy and regulate blood chemicals. But more than one in seven American adults have chronic kidney disease (CDC). To analyze the risk factors of chronic kidney disease is essential.

Our group use the data collected by Dr.P.Soundarapandian.M.D., D.M, a senior Consultant Nephrologist, who worked at Apollo hospitals. The response variable is a binary categorical variable, the logistic regression is one of the best suitable methods. And because there are a lot of features, 26, and there exist highly correlated between some variables depending on the medical literature review, thus, our group determine to use a random forest algorithm to fit the data, this method can avoid the multicollinearity issue in theory, because each node of each tree is constructed by finding single predictor and cutpoint for it (Deryło, L, 2017).

2. Data exploration

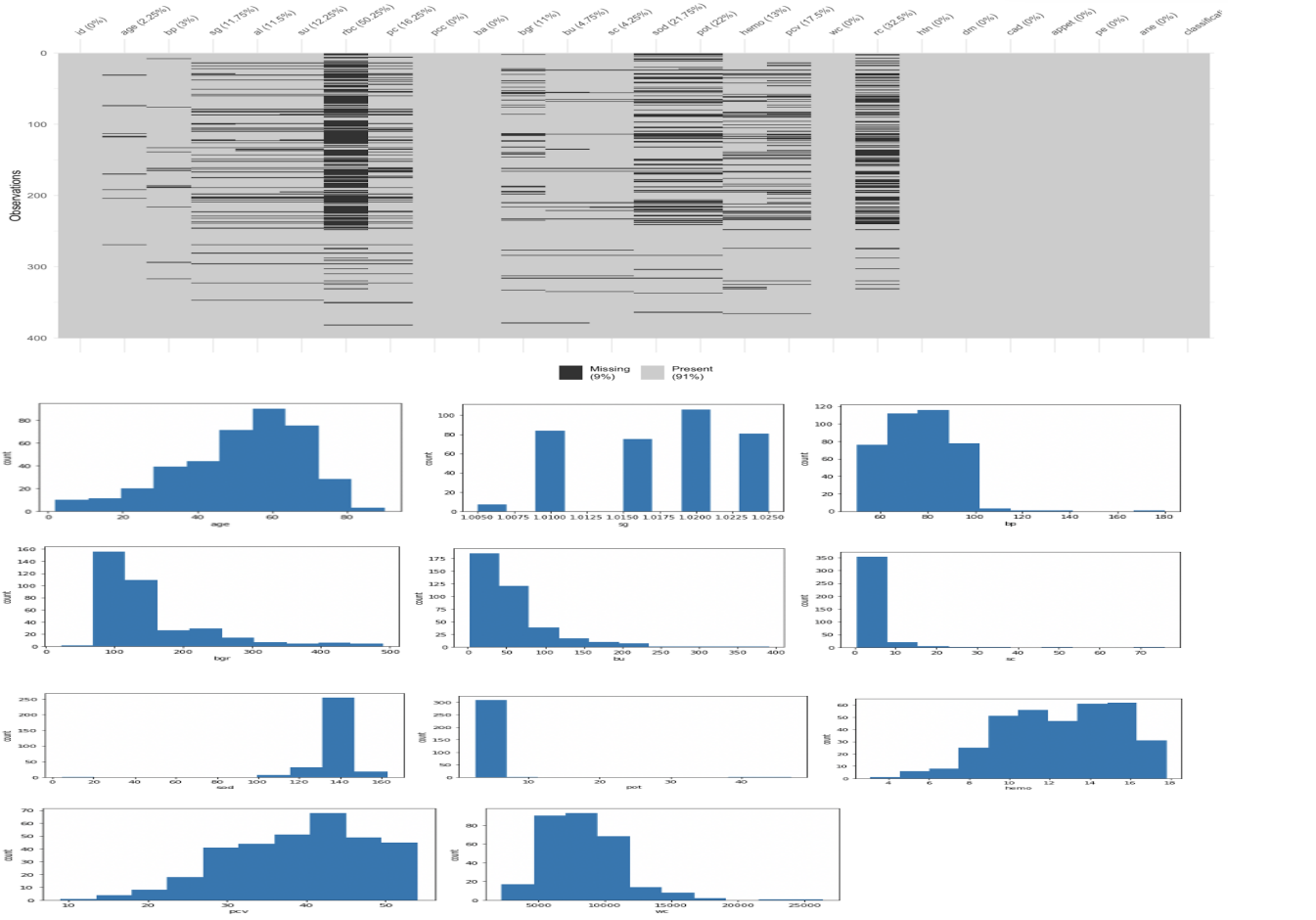
We see our dataset contains 400 observations with 26 variables. Our response variable is whether this person has chronic kidney disease. This is a classification problem, the response variable has two levels, respectively *'ckd','notckd',* which means whether this person has chronic kidney disease or not. Our response variable is basically balanced, the ratio is 5:3, so we don't need to resample the dataset.

We have some basic findings through table categorical variable with the response variable, for some continuous variables, such as age, we grouped age, corresponding to the response variable, and we find that people aged higher than 40 may be more likely to have chronic kidney disease. And people with abnormal red blood cells and pus cells have chronic kidney disease. We find some very strong variables that can identify whether people have chronic kidney disease, which are bacteria, hypertension and diabetes mellitus, and coronary artery disease. All people with the above diseases also have chronic kidney disease.

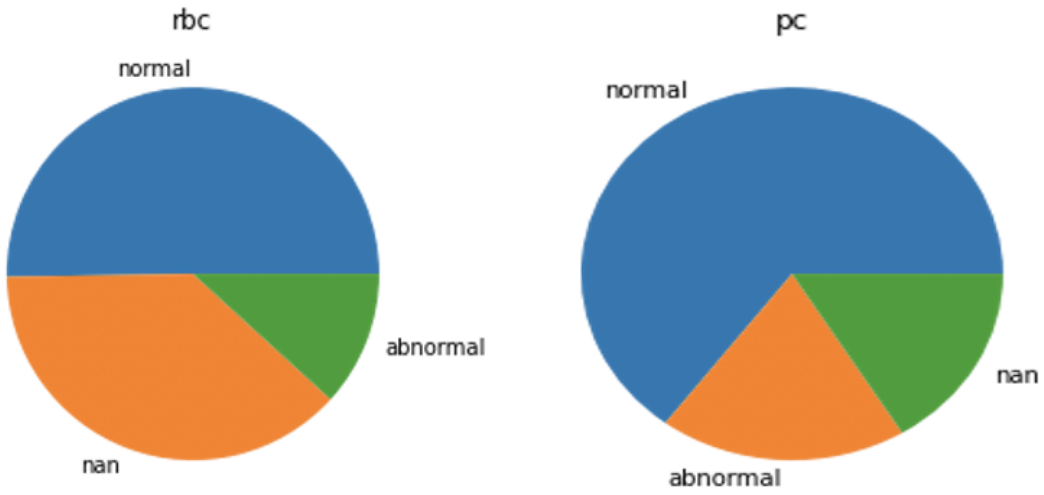
People with poor appetite, and pedal edema, anemia also have chronic kidney disease.

Then, we check missing values in our dataset and get the information for categorical variables and continuous variables. Our dataset contains so many NA and empty strings as missing values. From plot 1, we can see how large the missing value part is for each variable. I first turn all missing values into NA. And for numeric variables, we impute their missing values by mean or median and mode depending on their distribution plot, if their distribution is symmetric, we impute them by mean, if the distribution concentrates on the left side of the mean, we use the median to fill the missing value. We can see the variable '*rbc*' missing over 50%, '*rc'* with 32% missing values, *'sod*' over 21%, '*pcv*' has 17% missing value, '*hemo*' and '*bgr*', '*al*', '*su*', '*al*', '*sg*' are all past 10% missing value.

```
 [1] "age"  "bp"   "sg"   "al"   "su"   "rbc"  "pc"   "bgr"  "bu"   "sc"   "sod"  "pot"
[13] "hemo" "pcv"  "rc"
[1] "contain empty string:"
[1] "pcc"   "ba"    "wc"    "htn"   "dm"    "cad"   "appet" "pe"    "ane"
```



For categorical variables, we choose to use random forest models to impute the missing value.



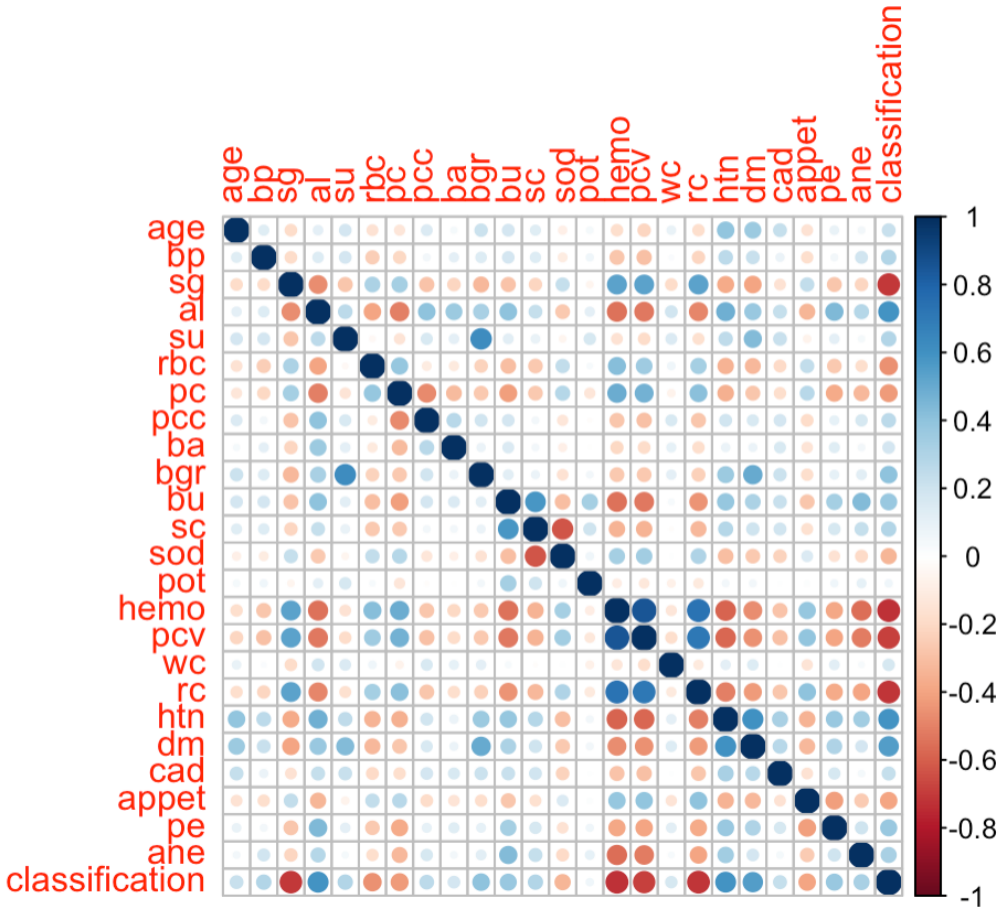
3. Data analysis

- Literature review

We do a correlation matrix and analyze it combined with medical knowledge. We can see that there are many factors highly correlated with each other. Specifically, *'rc', 'hemo', 'pcv*', and 'ane' are highly correlated because the first three factors are all measurements of blood cells and are risk factors for anemia which is 'ane'. The '*htn'* variable, which is hypertension is also correlated with '*rc', 'hemo*', and '*pcv'* since they are also risk factors for hypertension. It is also noticed that 'sod' is highly correlated with 'sc' as one study showed high sodium intake was associated with a greater decline in creatinine. There are also other correlated factors such as 'pc' and '*pcc*' which are pus cell-related factors, 'bu' and 'sc' which are all the nitrogenous end products of metabolism, and so on, they all have biological relevant reasons.

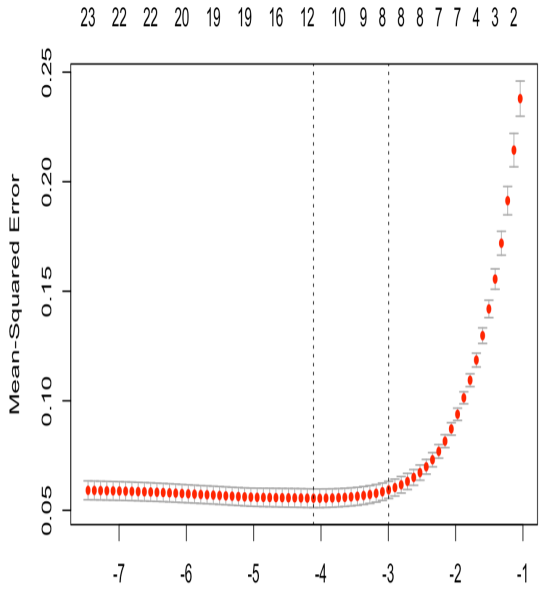
- Feature Selection

This is the most important step of our classification learning. We first draw a correlation matrix to research whether there exist highly correlated features.



From the correlation plot, we can see 'sg', '*rc', 'hemo*' and '*pcv*' are highly correlated with our response variable. And there also exist high correlations among other factors. In order to avoid the multicollinearity issue, which will result in overfitting, improve the standard error with the lower significant feature, and affect the finding of significant features of the response variable. We use lasso regression to help us filter the features. Lasso regression tries to minimize the cost function, which is an automatic regularization method, it will select useful features, and shrink the coefficient for those redundant features to 0.

We randomly split our data into a training set and a test set with a proportion of 7:3, fitting a lasso logistic regression model. We plot the mean square error of models under various values of λ and we notice that when log(*λ*) takes a value of approximately 0.01, the model minimizes the deviance. Thus, we fit the lasso logistic regression model with this λ value, and we obtain the result as below.



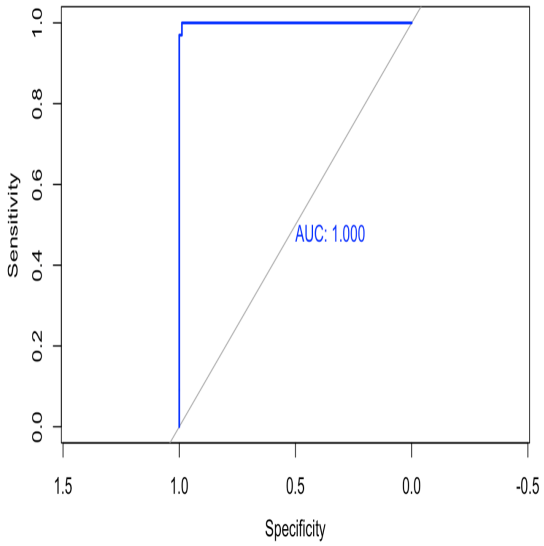
	s0
(Intercept)	6.961020e+02
age	.
bp	3.960016e-02
sg	-6.613139e+02
al	1.529922e+00
su	.
rbc	-3.915262e-01
pc	-2.026137e-01
pcc	.
ba	.
bgr	2.573649e-02
bu	6.089387e-03
sc	2.913305e-01
sod	-9.651238e-03
pot	.
hemo	-1.000938e+00
pcv	-1.841371e-01
wc	.
rc	-7.800197e-01
htn	7.600119e-01
dm	5.189084e+00
cad	.
appet	-4.100901e+00
pe	.
ane	.

We notice that 9 parameters are shrunk to 0, those are insignificant effects in lasso regression. The specific gravity, albumin, blood glucose random, hemoglobin, red blood cell count in millions per cmm, hypertension, diabetes mellitus, applet, red blood cells, packed cell volume are significant effect on whether this person has chronic kidney disease. The equation we get from this output is:

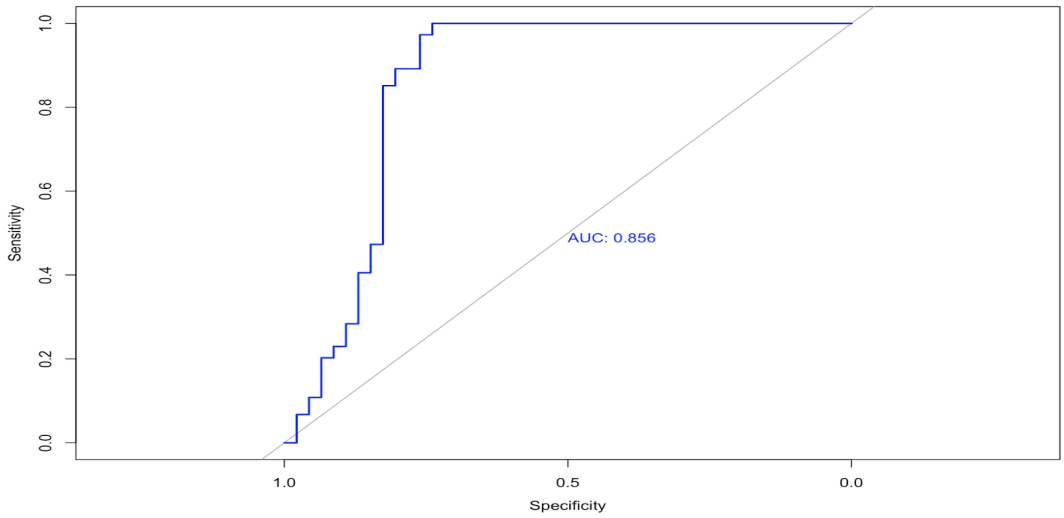
$$\log\left(\frac{p(ckd)}{p(not\ ckd)}\right) = 6.9e^{2} + 3.96e^{-2}\,bp - 6.61e^{2}sg + 1.53al - 3.92e^{-1}rbc - 2.03e^{-1}pc + 2.57e^{-2}bgr - 6.08e^{-3}bu + 2.91e^{-1}\text{sc} - 9.65e^{-3}sod - \text{hemo} - 1.84e^{-1}pcv - 7.8e^{-1}\,rc + 7.6e^{-1}\,htn + 5.18dm - 4.1appet$$

According to the function, we know the *'bp', 'al', 'bgr', 'sc', 'dm', 'htn'* are positive relations to the probability of people having chronic kidney disease. They canbe treated as riaks factors of chronic kidney disease. We test the predictive accuracy in test dataset, get 99% accuracy. The true positive rate is 100%.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 33  0
         1  1 86

               Accuracy : 0.9917
                 95% CI : (0.9544, 0.9998)
    No Information Rate : 0.7167
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9793

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9706
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9885
             Prevalence : 0.2833
         Detection Rate : 0.2750
   Detection Prevalence : 0.2750
      Balanced Accuracy : 0.9853
```



After fitting lasso logistic regression for the full model, I plug these filter variables as predictors to fit lasso logistic again, achieving 90% predictive accuracy. The true positive rate is 0.86. And we find the parameters of variables *'bp', 'sg', 'al', 'rbc', 'pc', 'bgr', 'ba', 'sc', 'sod', 'hemo', 'pcv', 'rc', 'htm', 'dm', 'appet'* are kept. These are significant factors for predicting chronic kidney disease.

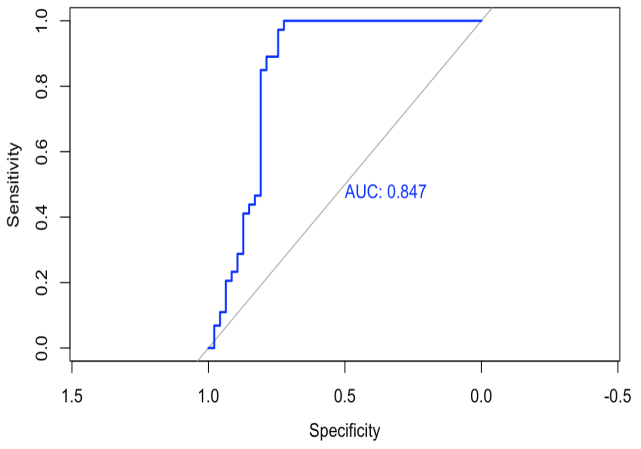


```
                        s0
(Intercept)  6.961020e+02
age           .
bp            3.960016e-02
sg           -6.613139e+02
al            1.529922e+00
su            .
rbc          -3.915262e-01
pc           -2.026137e-01
pcc           .
ba            .
bgr           2.573649e-02
bu            6.089387e-03
sc            2.913305e-01
sod          -9.651238e-03
pot           .
hemo         -1.000938e+00
pcv          -1.841371e-01
wc            .
rc           -7.800197e-01
htn           7.600119e-01
dm            5.189084e+00
cad           .
appet        -4.100901e+00
pe            .
ane           .
```

```
          Reference
Prediction  0  1
         0 34 12
         1  0 74

               Accuracy : 0.9
                 95% CI : (0.8318, 0.9473)
    No Information Rate : 0.7167
    P-Value [Acc > NIR] : 9.145e-07

                  Kappa : 0.7775

 Mcnemar's Test P-Value : 0.001496

            Sensitivity : 1.0000
            Specificity : 0.8605
         Pos Pred Value : 0.7391
         Neg Pred Value : 1.0000
             Prevalence : 0.2833
         Detection Rate : 0.2833
   Detection Prevalence : 0.3833
      Balanced Accuracy : 0.9302

       'Positive' Class : 0
```

And when I fit the lasso logistic regression model containing variables *'sg', 'hemo', 'pcv', and 'rc'* only, the predictive accuracy is also achieved at 90%, ture positive rate achieves 85%, it's not bad. Since we think these four variables are the most important effect on chronic kidney disease.

```
          Reference
Prediction  0  1
         0 34 13
         1  0 73

               Accuracy : 0.8917
                 95% CI : (0.8219, 0.941)
    No Information Rate : 0.7167
    P-Value [Acc > NIR] : 3.109e-06

                  Kappa : 0.7609

 Mcnemar's Test P-Value : 0.0008741

            Sensitivity : 1.0000
            Specificity : 0.8488
         Pos Pred Value : 0.7234
         Neg Pred Value : 1.0000
             Prevalence : 0.2833
         Detection Rate : 0.2833
   Detection Prevalence : 0.3917
      Balanced Accuracy : 0.9244
```



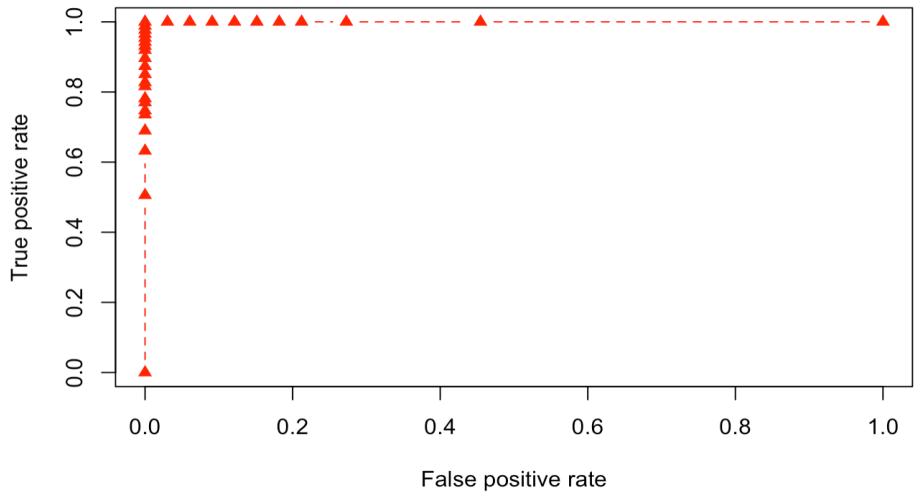
- Random Forest

Other than the lasso logistic regression model, we use the random forest model to fit our data, this method shows the importance of each factor in terms of the Mean Decrease Accuracy. Mean Decrease Accuracy(MDA) is the accuracy decreasing when we remove this one variable. Besides this, random forest algorithms can effectively avoid overfitting, it's probability-based. When I set up 100 trees, each tree will give us a prediction of chronic kidney disease whether this person has chronic kidney disease or not. This algorithm will return the prediction with a higher total number. Such as, if there are more than 50 trees prediction result is 'ckd', our prediction would be this person has chronic kidney disease. From the plot, we can see the order of importance is '*hemo', 'pcv', 'sg', 'rc', 'sc'*. This finding is consistent with our statements obtained through the lasso logistic regression model.

The predictive accuracy we get is 99%, and the positive rate is 100%. we suspect this model is overfitted. But it's reasonable, our dataset only has 400 observations and 24 factors. We build 100 trees, each tree containing up to 25 nodes. And there exist factors that have strong relations with the response variable, it can separate the response variable perfectly.

```
       MeanDecreaseGini
age          1.21010658
bp           0.68597718
sg          20.86778011
al           9.39335482
su           0.55848861
rbc          2.22739869
pc           0.49493591
pcc          0.05469919
ba           0.02400000
bgr          4.48475080
bu           3.39939860
sc          11.40563254
sod          2.08347013
pot          0.56086945
hemo        23.83692121
pcv         21.14531715
wc           0.83042566
rc          20.13435431
htn          6.14919541
dm           5.21170476
cad          0.00000000
appet        0.16543628
pe           0.37901668
ane          0.23935622
[1] 0.9916667
Confusion Matrix and Statistics
```

```
          Reference
Prediction  0  1
         0 33  0
         1  1 86

               Accuracy : 0.9917
                 95% CI : (0.9544, 0.9998)
    No Information Rate : 0.7167
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9793

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9706
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9885
             Prevalence : 0.2833
         Detection Rate : 0.2750
   Detection Prevalence : 0.2750
      Balanced Accuracy : 0.9853

       'Positive' Class : 0
```



4. Conclusion

In this research, we find '*sg', 'rc', 'hemo'*, and '*pcv*' are the most significant factors in predicting whether people have chronic kidney disease. If there are significant changes in these indicators, people should pay attention.

The predictive accuracy of our random forest model is too high. And the dataset we use has too many missing values, for those numeric variables with over 15% missing values, I use mean or median to impute according to their distribution. I think it would more convincing if the data do not have so much missing.

Based on Occam's razor also known as the law of parsimony, we tend to choose a simpler model even though it sacrifices some accuracy. Our final model contains '*sg', 'rc', 'hemo', 'pcv'* reaches a pretty high accuracy and AUC. In addition, based on a study, '*pcv*' was found significantly low in the CKD group. Thus it is reasonable to use the model as our predicting model.

Reference

Hematocrit/packed cell volume. (n.d.). April 29, 2022. Retrieved from https://eclinpath.com/hematology/tests/hematocrit/

Smyth, A., O'Donnell, M. J., Yusuf, S., Clase, C. M., Teo, K. K., Canavan, M., . . . Mann, J. F. (2014). A systematic review. *American Journal of Hypertension*

Göbel BO;Schulte-Göbel A;Weisser B;Glänzer K;Vetter H;Düsing R;. (n.d.).April 29, 2022. Arterial blood pressure. correlation with erythrocyte count, hematocrit, and hemoglobin concentration. Retrieved from https://pubmed.ncbi.nlm.nih.gov/2006992/

Iseki, K., & Kohagura, K. (2007, November 01). Anemia as a risk factor for chronic kidney disease. Retrieved April 29, 2022, from https://www.sciencedirect.com/science/article/pii/S0085253815525568

Mamun, I. (2021, December 10). *Multicollinearity in Data Science - Towards Data Science*. Medium. https://towardsdatascience.com/multicollinearity-in-data-science-c5f6c0fe6edf

Yiu, T. (2021, December 10). *Understanding Random Forest - Towards Data Science*. Medium. https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:%7E:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree.

Maklin, C. (2021, December 11). *Random Forest In R - Towards Data Science*. Medium. https://towardsdatascience.com/random-forest-in-r-f66adf80ec9