08/09/2017

Predicting Early Stage of Chronic Kidney Disease Based on Blood Test Result

Tom Li

New York University

Table of Contents

[Phase 1: Business Understanding 3](#_Toc490088349)

[Phase 2: Data Understanding 4](#_Toc490088350)

[Attribute Information 4](#_Toc490088351)

[Acronyms in Dataset 6](#_Toc490088352)

[Class Distribution 7](#_Toc490088353)

[Phase 3: Data Preparation 8](#_Toc490088354)

[Filling Missing Values 8](#_Toc490088355)

[Data Reduction & Feature Selection. 11](#_Toc490088356)

[CFS Forward Selection: 11](#_Toc490088357)

[Information Gain 12](#_Toc490088358)

[Gain Ratio 12](#_Toc490088359)

[Phase 4: Predictive Modeling 16](#_Toc490088360)

[Cluster Analysis 16](#_Toc490088361)

[Classification 17](#_Toc490088362)

[Conclusions 18](#_Toc490088363)

# Phase 1: Business Understanding

Chronic kidney disease (CKD) is a common disease that affects billions of people around the globe. It is a condition characterized by a gradual loss of kidney function over time. Kidney is one of the most important organs in human body. If a kidney failure happens, it could be life-threatening because kidney is responsible for disposing wastes and balancing salt levels in our blood as well as stimulating the production of red blood cells. It is very important for doctors to be able to diagnose CKD in its early stages and to prescribe proper drugs that treat this condition. A simple blood test can have a lot of implications of the health of kidney and it is easily accessible and relatively cheap for public to get diagnose. So we decided to base our model on blood test result that can actually be implemented in real life considering its low cost and easy access. If successful the model built in the project could profit many pharmaceutical companies and most importantly, it can save millions of lives.

In order to discover our hypothesis and to understand how we are going to build a predicting model, we should first take a look at what causes CKD. Hypertension and diabetes are the two major causes of this disease. It would be safe to assume that patients with CKD are more likely to have higher blood pressure and higher sugar level. However, CKD might have other effects on blood if we think about what a kidney does. As mentioned before, kidney’s three main functions are filtering out wastes, maintaining salt levels and producing red blood cell stimulating hormone. So it is logical to think that a CKD patient will likely have more wastes, abnormal salt levels and less red blood cells in their blood.

# Phase 2: Data Understanding

The data used in this project was downloaded from UCI Machine Learning website: <https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease> and it includes 400 Indian patients’ blood test results. The dataset was created by L. Jerlin Rubini from Dr. P. Soundarapandian in 2015. There are 25 attributes in this dataset including a class attribute in the last column indicating if a patient is tested positive (ckd) or negative (notckd) for CKD. Details of the attributes are as follows:

## Attribute Information

1. Age(numerical)

age in years

1. Blood Pressure(numerical)

bp in mm/Hg

1. Specific Gravity(nominal)

sg - (1.005,1.010,1.015,1.020,1.025)

1. Albumin(nominal)

al - (0,1,2,3,4,5)

1. Sugar(nominal)

su - (0,1,2,3,4,5)

1. Red Blood Cells(nominal)

rbc - (normal,abnormal)

1. Pus Cell (nominal)

pc - (normal,abnormal)

1. Pus Cell clumps(nominal)

pcc - (present,notpresent)

9.Bacteria(nominal)

ba - (present,notpresent)

10.Blood Glucose Random(numerical)

bgr in mgs/dl

11.Blood Urea(numerical)

bu in mgs/dl

12.Serum Creatinine(numerical)

sc in mgs/dl

13.Sodium(numerical)

sod in mEq/L

14.Potassium(numerical)

pot in mEq/L

15.Hemoglobin(numerical)

hemo in gms

16.Packed Cell Volume(numerical)

17.White Blood Cell Count(numerical)

wc in cells/cumm

18.Red Blood Cell Count(numerical)

rc in millions/cmm

19.Hypertension(nominal)

htn - (yes,no)

20.Diabetes Mellitus(nominal)

dm - (yes,no)

21.Coronary Artery Disease(nominal)

cad - (yes,no)

22.Appetite(nominal)

appet - (good,poor)

23.Pedal Edema(nominal)

pe - (yes,no)

24.Anemia(nominal)

ane - (yes,no)

25.Class (nominal)

class - (ckd,notckd)

## Acronyms in Dataset

age - age

bp - blood pressure

sg - specific gravity

al - albumin

su - sugar

rbc - red blood cells

pc - pus cell

pcc - pus cell clumps

ba - bacteria

bgr - blood glucose random

bu - blood urea

sc - serum creatinine

sod - sodium

pot - potassium

hemo - hemoglobin

pcv - packed cell volume

wc - white blood cell count

rc - red blood cell count

htn - hypertension

dm - diabetes mellitus

cad - coronary artery disease

appet - appetite

pe - pedal edema

ane - anemia

class - class

## Class Distribution

ckd: 250

notckd: 150

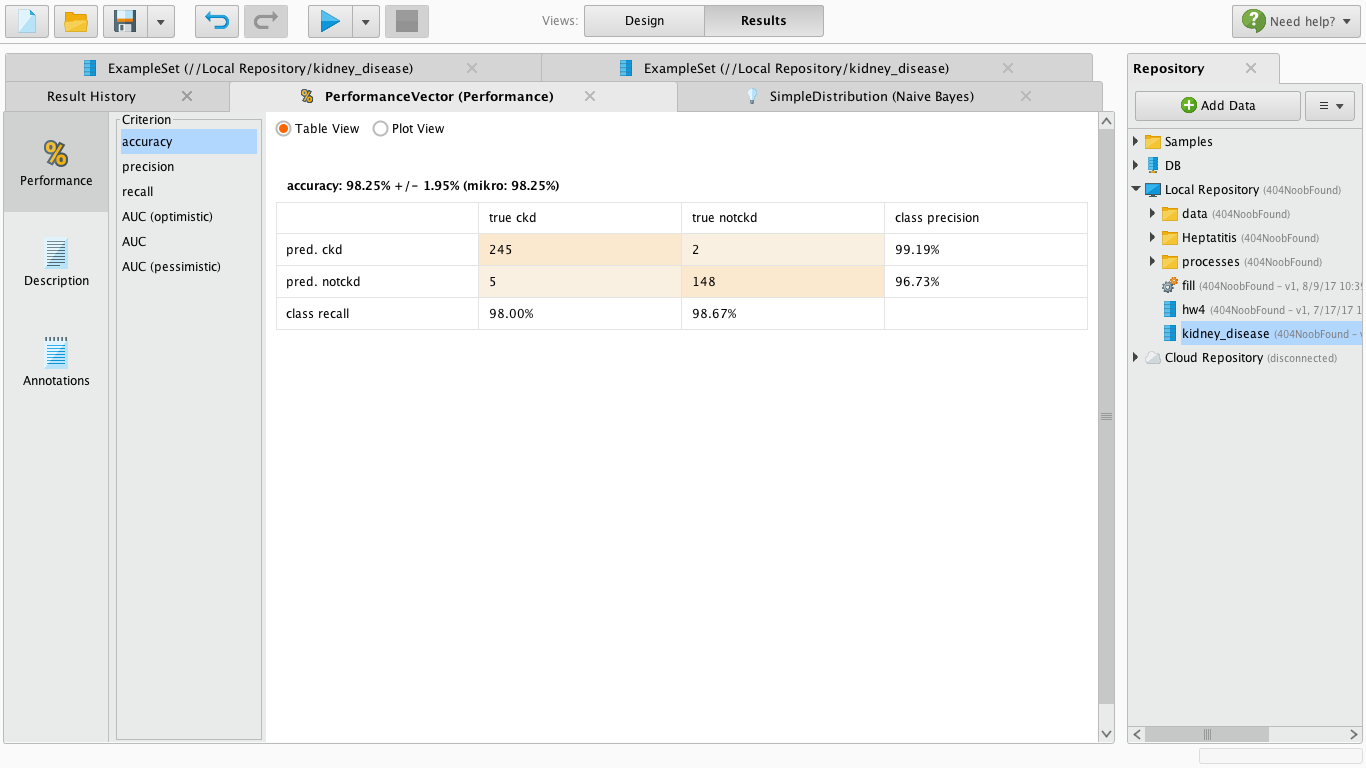
It seems that the data quality is good even though there are missing values of some attributes. Most of the attributes can be found on a regular blood test report and the rest can be easily collected from each patient.

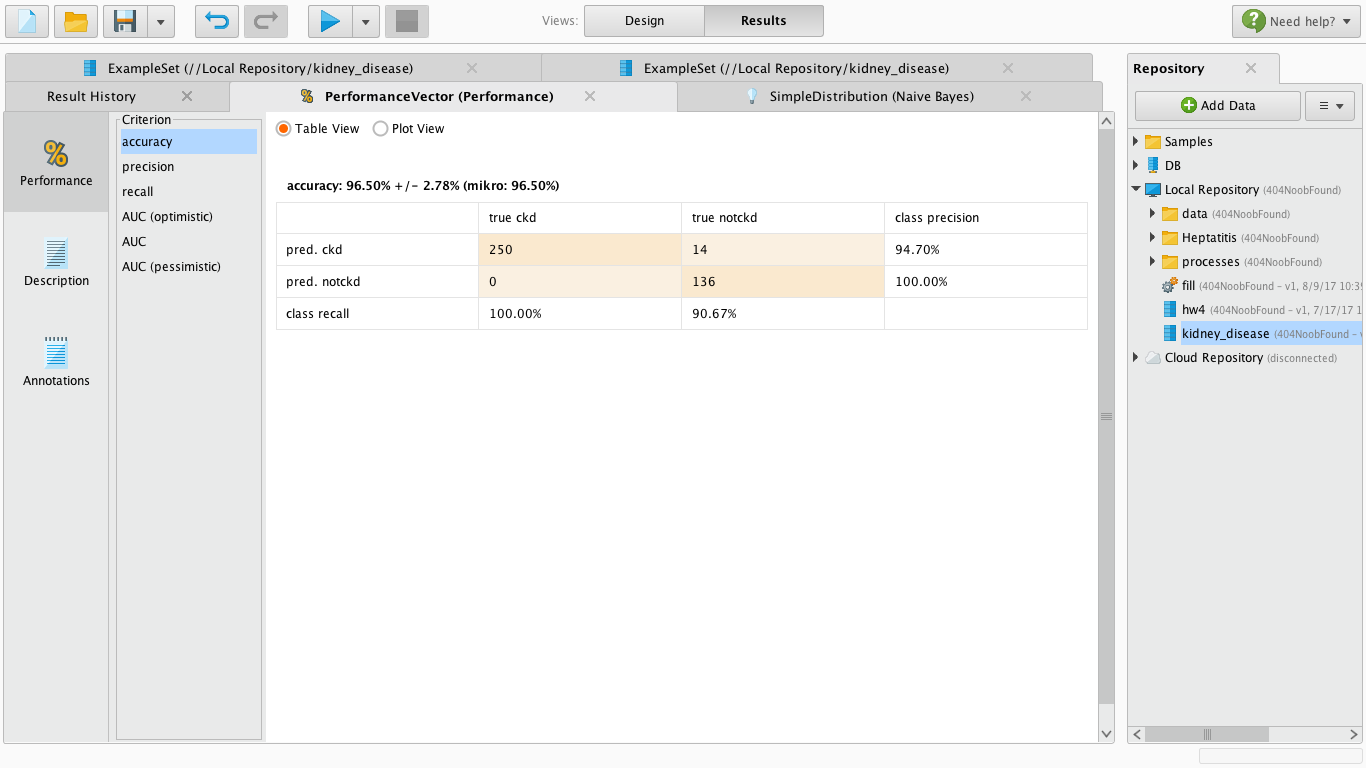
This project uses several different data analytics tools such as RapidMiner, R Studio and Weka. The tool used for each data mining process will be specified in the following phases.

# Phase 3: Data Preparation

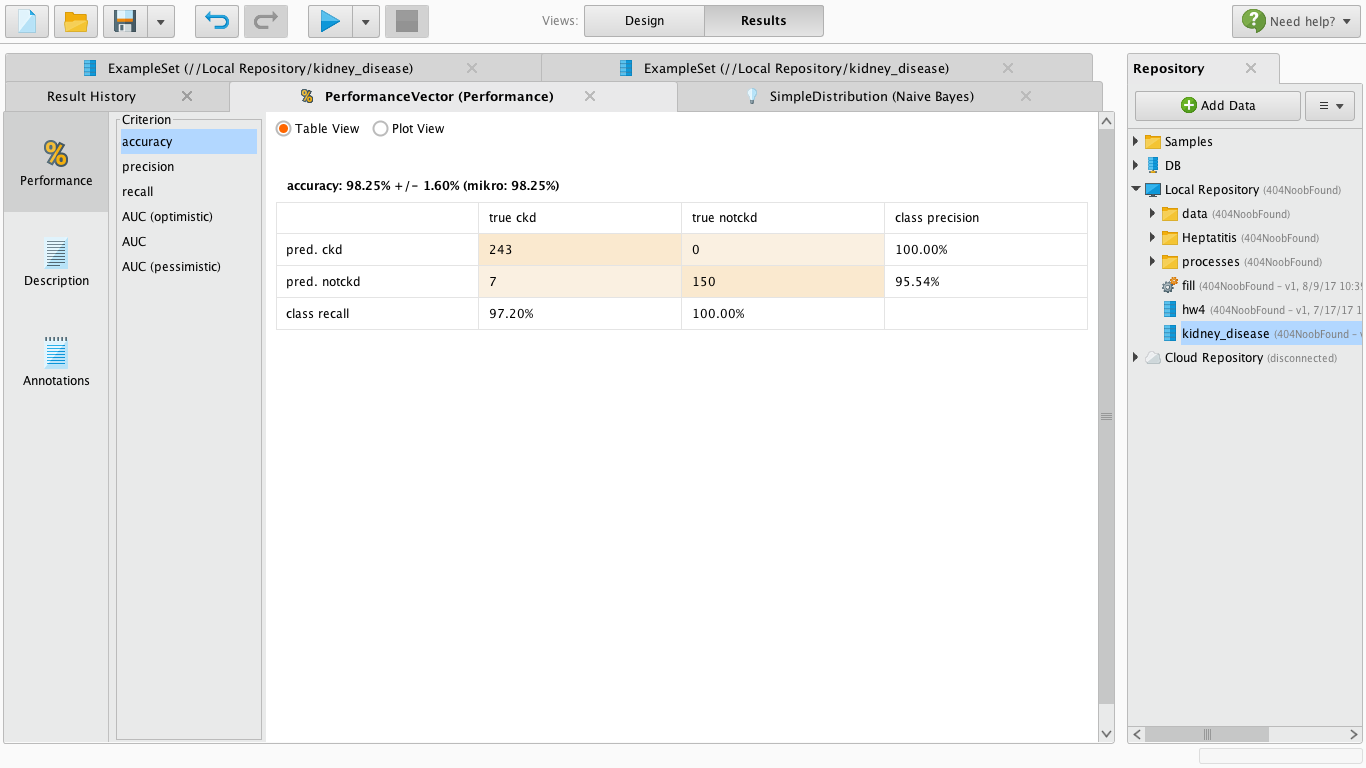
## Filling Missing Values

Since there are missing values, we will need to fill those first. The original data file is titled “kidney\_disease.csv” and was loaded it in RapidMiner for this task specifically because it provides different strategies for filling missing values: minimum, maximum, average, zero, none. In order to get the best result, we compared the accuracy of each strategy and they are listed as follows:

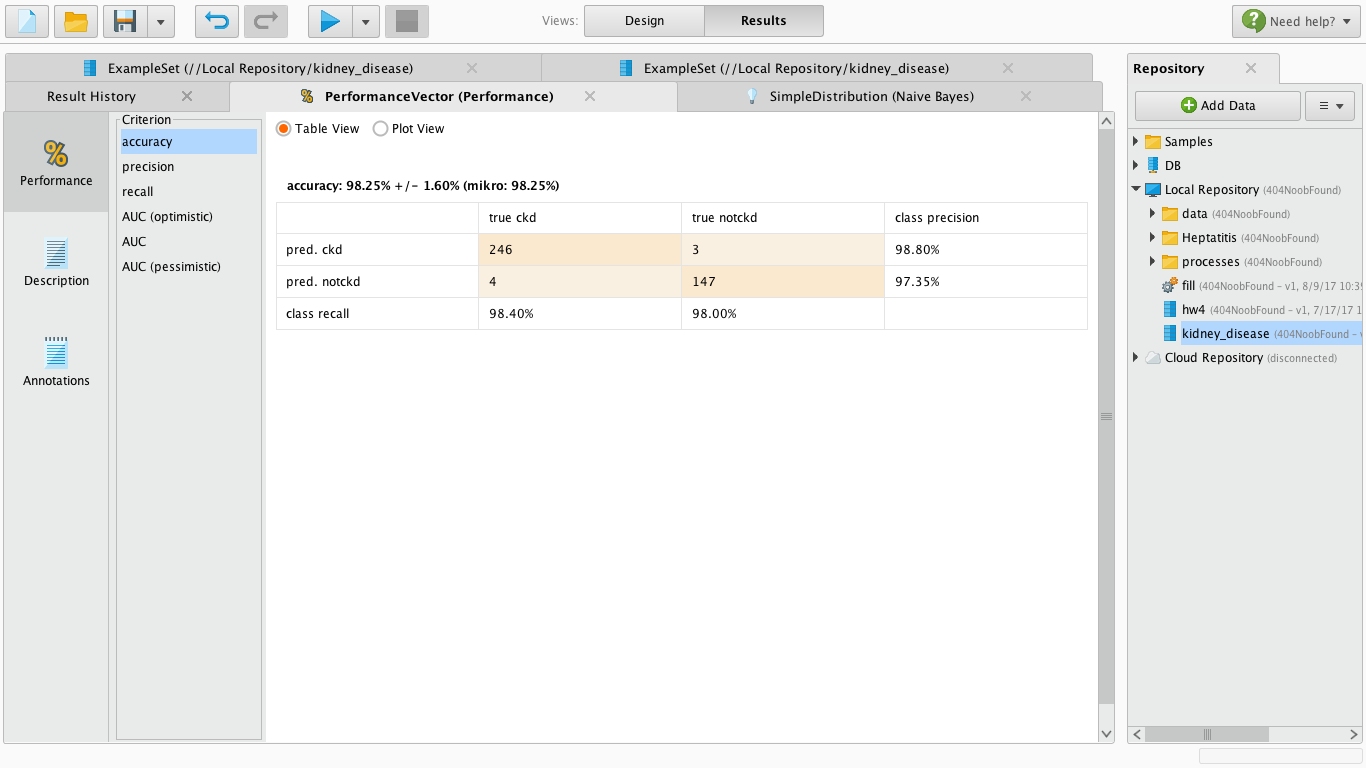
Replacing missing values with minimum. Accuracy = 98.25%.



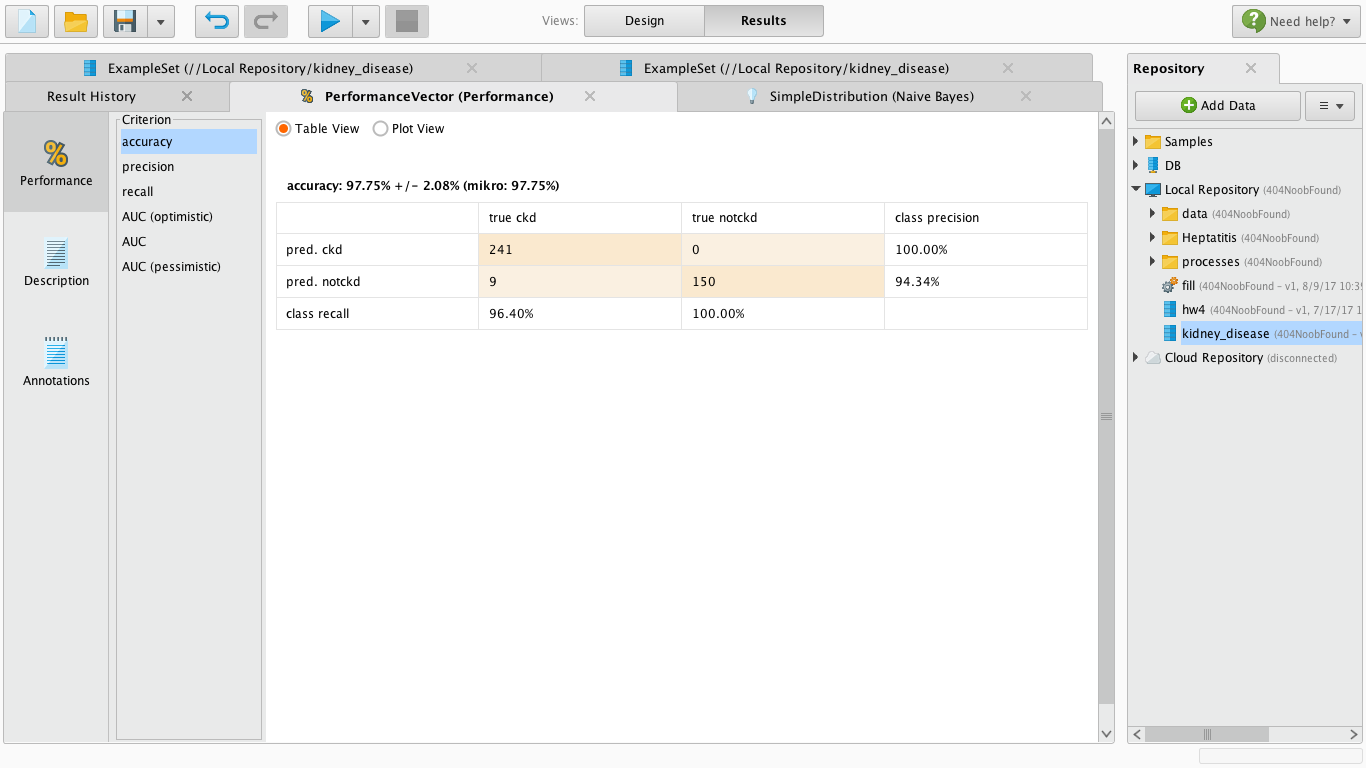
Replacing missing values with maximum. Accuracy = 96.50%.



Replacing missing values with average. Accuracy = 98.25%.



Replacing missing values with zero. Accuracy = 98.25%.



Replacing missing values with none. Accuracy = 97.75%.

It is obvious that the minimum, average and zero have equally highest accuracy. However by comparing the confusion matrix of the three, the average strategy is the best at predicting CKD (class prediction: 100%) meaning that if a patient is predicted as having CKD, then this patient is 100% having CKD. So replacing the missing values with the average of each attribute is clearly the better strategy here. We then output a new data file “fill.csv” that each cell has been filled entirely.

## Data Reduction & Feature Selection.

Data reduction and feature selection is an important step in prepare data for modeling. Because it reduces the size of the dataset by choosing only relevant attributes without compromising accuracy of our model analysis. Principle Component Analysis (PCA) was performed for dimensionality reduction (test results are shown in file “pca”) using Weka. It took the original dataset as a matrix and transformed it into another matrix filled by the Eigen vectors of each selected attribute. However, due to the drawbacks of PCA, it does not preserve the actual attributes and their values and since the interpretability of data is a key in this project, PCA was not preferred.

Nevertheless, we still need to get the “essence” of this dataset so we moved on to feature selection where a subset of features is selected from all 25 attributes. Again, there are multiple ways to achieve this task, such as entropy-based filters, Chi-squared filter, Correlation Feature Selection (CFS), etc. We performed some of these in Weka and here are the results.

### CFS Forward Selection:

CFS forward selection is a greedy algorithm that starts with an empty set and sequentially add the feature x+ that results in the highest value of the objective function J(Yk+x+) when combined with the features Yk that have already been selected. Detailed result can be found in file “forward selection”. Here is a preview:

Attributes selected: bp, sg, al, rbc, bgr, sc, sod, pcv, rc, htn, dm, appet, pe, ane

Total: 15

### Information Gain

Information gain is an entropy-based filter that evaluates the worth of an attribute by measuring the information gain with respect to the class. Detailed result can be found in file “infogain”. Here is a preview:

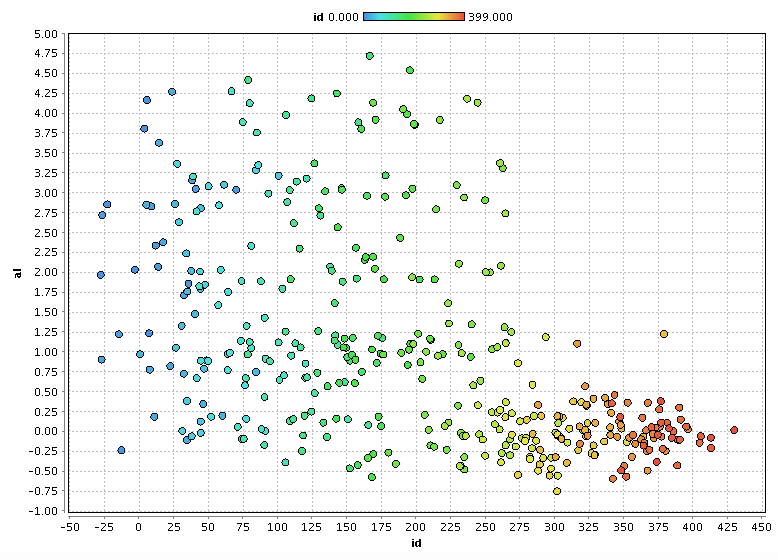
Attribute ranking from high to low: hemo, pcv, rc, sg, sc, al, sod, htn, pot, dm, bu, bgr, wc, bp, appet, pe, pc, su, ane, age, rbc, pcc, cad, ba

### Gain Ratio

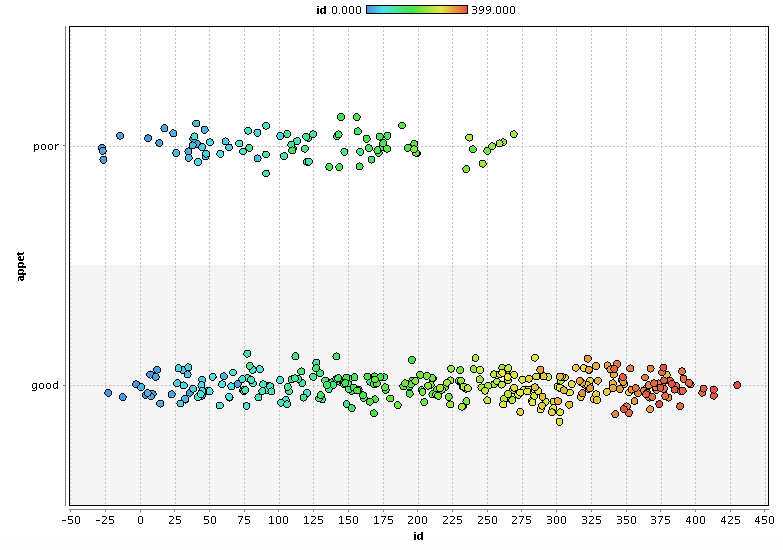
Gain ratio is also entropy-based filter and is a modification of information gain that reduces its bias. Gain ratio overcomes the problem with information gain by taking into account the number of branches that would result before making the split. It corrects information gain by taking the intrinsic information of a split into account. Detailed result can be found in file “gain ratio”. Here is a preview:

Attribute ranking from high to low: sc, pcv, hemo, htn, al, dm, sg, rc, bgr, bp, appet, bu, pc, pe, su, ane, rbc, pcc, sod, cad, pot, ba, wc, age

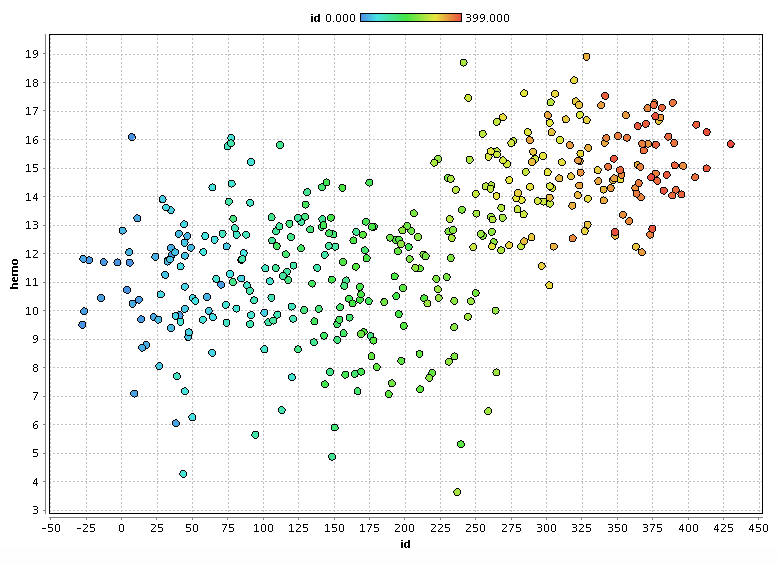
By comparing the results, a subset of 18 attributes was selected which includes bp, sg, al, su, pc, bgr, bu, sc, sod, pot, hemo, pcv, rc, htn, dm, appet, pe, ane. A new dataset was generated that only has these selected attributes and can be found in file “reduced”. Below are some visualizations of some selected attributes.



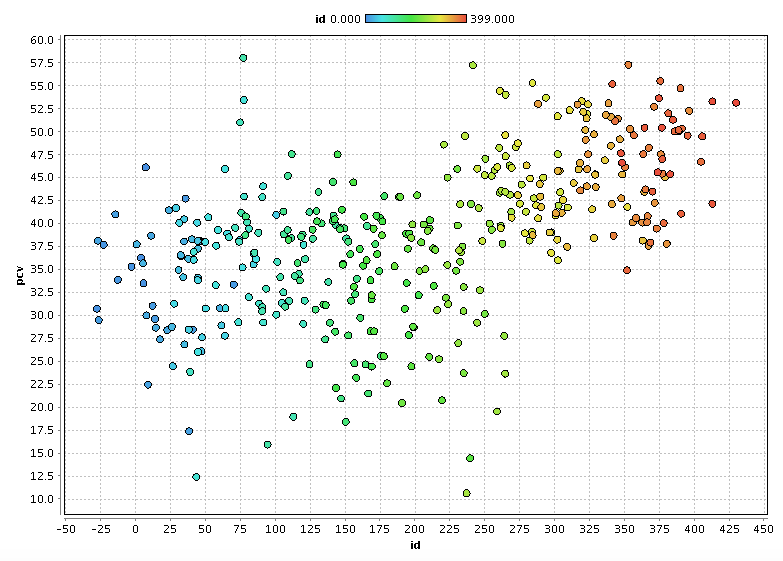
patient id vs al



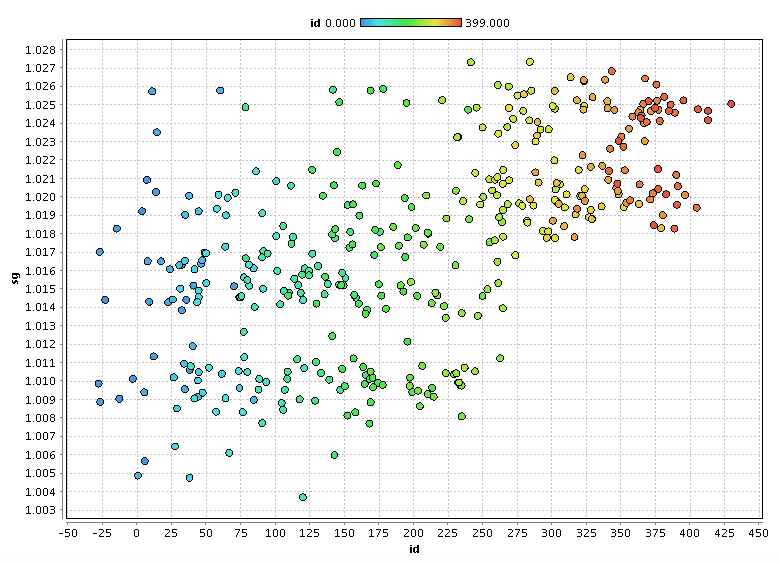
patient id vs appet



patient id vs hemo



patient id vs pcv



patient id vs sg

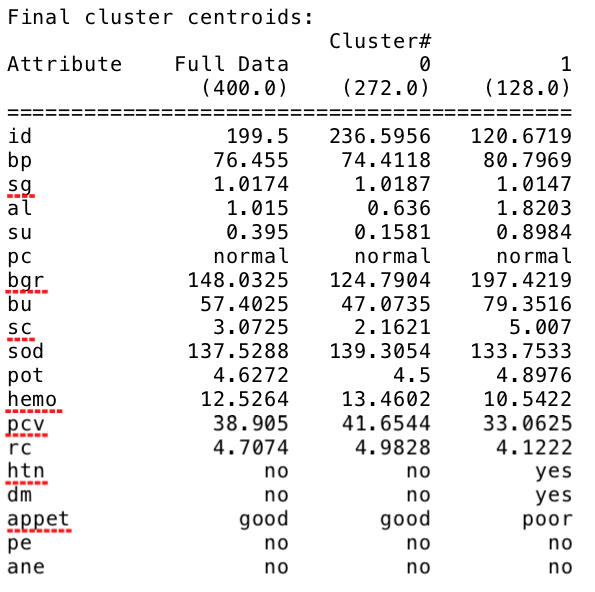
As we can see in these scatter plots, since we know that patients with id 0-249 have CKD and those with id 250-399 does not have CKD, it is not hard to see some correlation between each of these attributes and the class. For example, in the last graph patient id vs hemo, it seems that CKD patients have lower levels in hemo and patients without CKD have higher levels. More detailed explanation of these correlations will be discussed in the Modeling section. Now we have finished preparing our dataset and it is ready for modeling.

# Phase 4: Predictive Modeling

This project is a data classification project that predicts if a patient has CKD or not when given their blood test result. However, before building a classifier, it is helpful to perform a cluster analysis which can give us a general idea or what to expect for data classification.

## Cluster Analysis

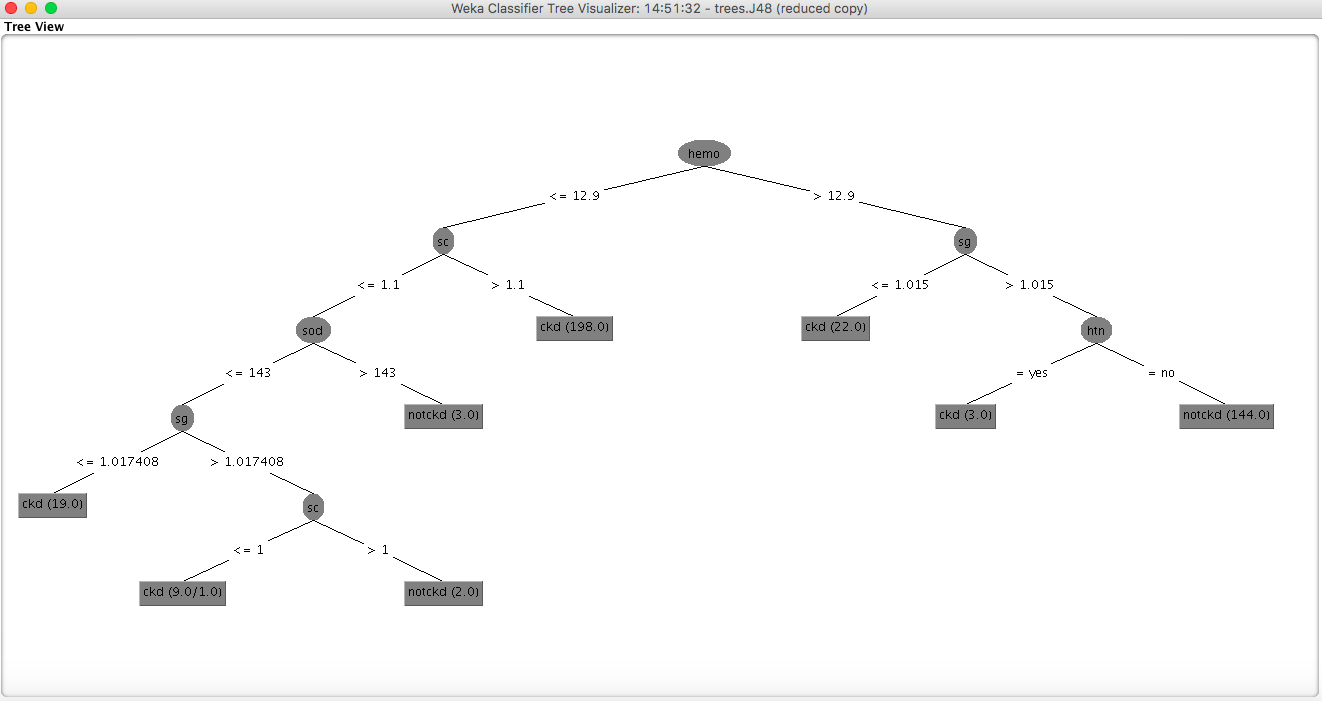
Data clustering is performed in Weka using kMeans++ algorithm. Two clusters were formed where each cluster was assigned to a class. Cluster#0 is notckd and Cluster#1 is ckd. The accuracy of this clustering is about 70% and detailed result can be found in file “cluster analysis Weka”. Here are the final centroids:



We can see here the result here can somewhat indicate the correlation between the each of the attributes and the class and corresponds to the scatter plots shown in the previous section. For example, for pcv, cluster#0 is higher than cluster#1, which is similar to the id vs pcv scatter plot where notckd patients are slightly higher than ckd patients. Also it should be noticed that htn is no for cluster#0 and yes for cluster#1 which is exactly our hypothesis in the first phase where we assumed that high blood pressure is a symptom of CKD.

## Classification

To build our classifier, a decision tree algorithm or J48 was performed using Weka. J48 is an open source java implementation of C4.5 algorithm which is an extension of ID3 algorithm. Detailed result can be found in file “classification weka”. Here is a visualization of the tree:



Surprisingly, this classifier has an accuracy of 99.25%. Only one instance was incorrectly classified. To interpret this tree, let’s recall what these acronyms stand for and what they mean:

hemo: hemoglobin, the protein molecule in red blood cells that carries oxygen from the lungs to the body's tissues and returns carbon dioxide from the tissues back to the lungs;

sc: serum creatinine, a waste product in your blood that comes from muscle activity;

sg: specific gravity, a urinalysis parameter commonly used in the evaluation of kidney function;

sod: sodium, salt;

htn: hypertension, high blood pressure;

## Conclusions

Check each of the following in its specific attribute order to determine if a new patient has CDK:

1. Low hemo (red blood cell), low sc (waste), low sodium and low sg: ckd
2. Low hemo (red blood cell), low sc (waste), low sodium, high sg and low sc (waste): ckd
3. Low hemo (red blood cell) and high sc (waste): ckd
4. High hemo (red blood cell) and low sg: ckd
5. High hemo (red blood cell), low sg and htn (high blood pressure): ckd

All of these attributes are associated with either the cause of CKD or the function of CKD and are in accordance with the assumptions we made in Phase 1.