Introduction

The kidneys are two bean-shaped organs, each about the size of a fist. They are located just below the rib cage, one on each side of the spine. Every day, the kidneys filter about 120 to 150 quarts of blood to produce about 1 to 2 quarts of urine. the function of the kidneys is to remove waste products and excess fluid from the body through the urine. The production of urine involves highly complex steps of excretion and re-absorption. This process is necessary to maintain a stable balance of body chemicals. The critical regulation of the body's salt, potassium and acid content is performed by the kidneys and produce hormones that affect the function of other organs. For example, a hormone produced by the kidneys stimulates red blood cell production, regulate blood pressure and control calcium metabolism etc.

Chronic kidney disease (CKD) is a major issue worldwide which is a condition characterized by a gradual loss of kidney function over time, 14% of the world population suffer from CKD. Over 2 million people worldwide currently receive treatment with dialysis or a kidney transplant to stay alive, yet this number may only represent 10% of people who actually need treatment to live. Chronic kidney disease, causes more deaths than breast cancer or prostate cancer and It is the under-recognized public health crisis.

The stages of CKD are mainly based in the measured or estimated Glomerular Filtration Rate (eGFR) which is based on Creatinine level, Gender, Race and Age. There are Five stages, kidney functionality if normal in stage 1 and minimally reduced in stage 2, but the majority of cases are stage 3.

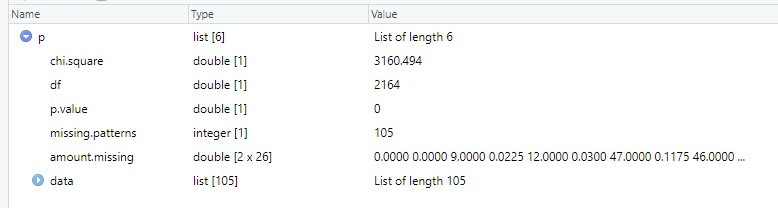
< Have to add the Machine Learning part and the details about the data set >

Preprocessing analysis of the Data set

The attributes with more than 20% missing values ("red blood cells", "sodium", "potassium", "white blood cell count" , "red blood cell count" ) has been discarded because of the error that will occurs when filling them.

|  |  |
| --- | --- |
| Atribute | missing pesentage |
| class | 0.00 |
| appetite | 0.25 |
| pedal edema | 0.25 |
| anemia | 0.25 |
| hypertension | 0.50 |
| diabetes mellitus | 0.50 |
| coronary artery disease | 0.50 |
| pus cell clumps | 1.00 |
| bacteria | 1.00 |
| age | 2.25 |
| blood pressure | 3.00 |
| serum creatinine | 4.25 |
| blood urea | 4.75 |
| blood glucose random | 11.00 |
| albumin | 11.50 |
| specific gravity | 11.75 |
| sugar | 12.25 |
| hemoglobin | 13.00 |
| pus cell | 16.25 |
| packed cell volume | 17.50 |
| sodium | 21.75 |
| potassium | 22.00 |
| white blood cell count | 26.25 |
| red blood cell count | 32.50 |
| red blood cells | 38.00 |

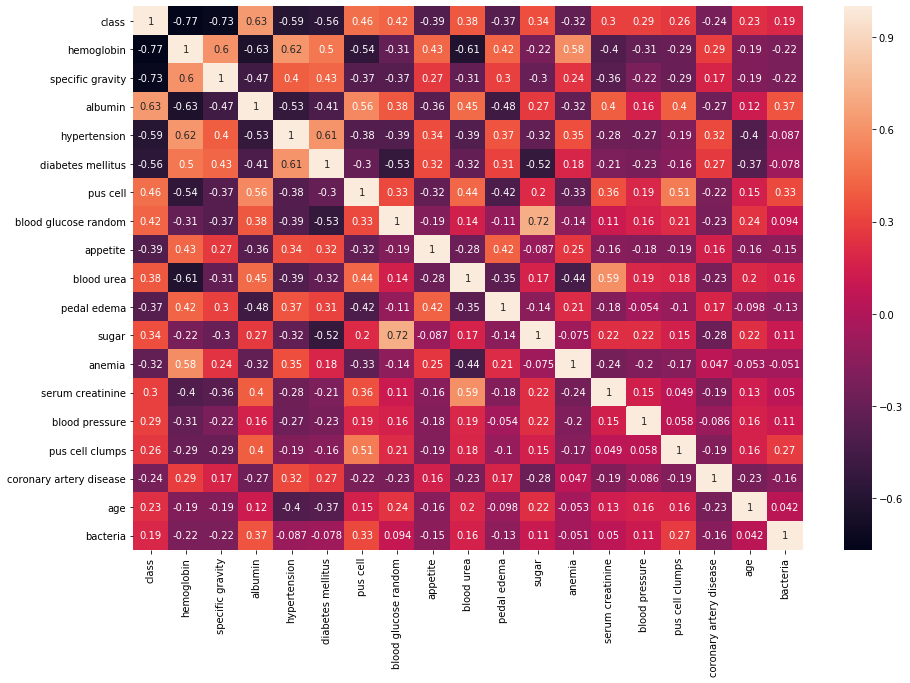
In the pre-processing the incomplete data handling grabs a major portion of it, depending on the way the missing data are distributed it should be treated in different ways with different assumption about them, one of them is missing completely at random, to validate this assumption the Little’s MCAR test has been done.

*If missing data mechanism depends on the unobserved data, data are missing not at random so chi-square test of MCAR for multivariate quantitative data proposed by Little (1988), which tests whether there exists significant difference between the 1 means of different missing-value patterns.*

From that it clearly shows from the “P” value is zero, so the data has missed completely randomly. Therefore, substituting missing values with a constant will rapidly drop the accuracy and it will bias the prediction since more CKD positive instances are there. Which led to fill missing values with a more algorithmic approach from K Nearest Neighbor Algorithm.

When applying the algorithm, since all patients are different from each other, which led to use number of estimators as same the number of complete instances, which gave the minimum mean and standard deviation change, since the missing values are missed completely randomly.

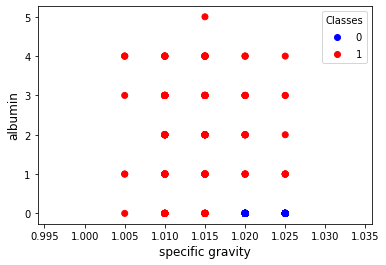
There after that the Pearson’s co-relation matrix has taken



When consider the absolute values of the above matrix of the class attribute. It clearly shows “Hemoglobin”, “Specific Gravity”, “Albumin”, “Hypertension” and “Diabetes mellitus” have the highest correlations more than 0.5. Then the secondary attributes “Pus cell”, “Blood glucose Random”, “appetite”, “Blood Urea”, “pedal edema”, “Sugar”, “Anemia” and “Serum Creatinine” are the attributes which has correlation more than 0.3.

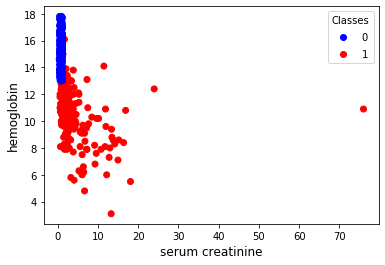
Specific Gravity and Albumin has only 5 sets of values in each, when plot the values in a 2D graph it clearly shows a cluster in that with CKD negative instances.

A test for protein in the urine. **Estimates the amount of a albumin that is in your urine**. An excess amount of protein in your urine may mean your kidney's filtering units have been damaged by disease or due to fever or heavy exercise, test should be confirmed over several weeks.



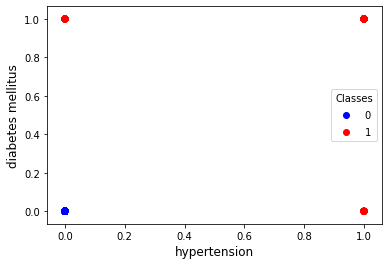
There are a number of ways to increase hemoglobin levels. In general, low hemoglobin levels that need to be increased are caused by three circumstances: decreased red blood cell production, increased red blood cell destruction and by blood loss. Healthy kidneys produce a hormone called erythropoietin (EPO). A hormone is a chemical produced by the body and released into the blood to help trigger or regulate particular body functions. EPO prompts the bone marrow to make red blood cells, which then carry oxygen throughout the body. When kidneys are diseased or damaged, they do not make enough EPO. As a result, the bone marrow makes fewer red blood cells, causing anemia but before it causes anemia (which happens after less than 50% of one kidney is properly functions), the hemoglobin levels changes slightly.

From the Hemoglobin vs Serum Creatinine 2D plot clearly shows 2 different clusters of CKD positive and Negative.



Serum Creatinine also known as Creat, Blood Creatinine, Creatinine. Creatinine is a waste product produced by muscles from the breakdown of a compound called creatine. Creatinine is removed from the body by the kidneys, which filter almost all of it from the blood and release it into the urine. This test measures the amount of creatinine in the blood. Creatine is part of the cycle that produces energy needed to contract muscles. Both creatine and creatinine are produced by the body at a relatively constant rate. Apart from issues directly related to kidney, a high-protein diet, congestive heart failure, complications of diabetes and dehydration can also increase the level of creatinine in the blood. The normal serum creatinine range is 0.6–1.1 mg/dL in women and 0.7–1.3 mg/dL in men.

The two main causes of chronic kidney disease are diabetes and high blood pressure, which are responsible for up to two-thirds of the cases. Diabetes happens when the blood sugar is too high, causing damage to many organs in the body, including the kidneys and heart, as well as blood vessels, nerves and eyes. High blood pressure, or hypertension, occurs when the pressure of your blood against the walls of your blood vessels increases. If uncontrolled, or poorly controlled, high blood pressure can be a leading cause of heart attacks, strokes and chronic kidney disease. Also, chronic kidney disease can cause high blood pressure.

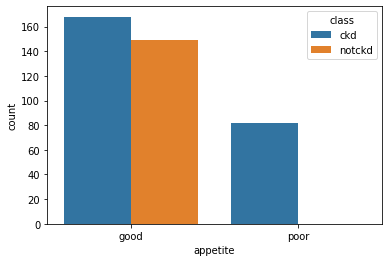


|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | hemoglobin | specific gravity | albumin | hypertension | diabetes mellitus | pus cell | blood glucose random | appetite | blood urea | pedal edema | sugar | anemia | serum creatinine |
| count | 13.00% | 11.75% | 11.50% | 0.50% | 0.50% | 16.25% | 11.00% | 0.25% | 4.75% | 0.25% | 12.25% | 0.25% | 4.25% |
| mean | 0.66% | 0.01% | -1.39% | -0.34% | -0.36% | -2.22% | -0.14% | 0.04% | -0.22% | -0.10% | -4.53% | -0.15% | -0.43% |
| std | -6.32% | -6.15% | -5.95% | -0.19% | -0.19% | -8.94% | -5.86% | -0.12% | -2.43% | -0.12% | -6.30% | -0.12% | -2.16% |
| min | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 25% | 5.29% | 0.49% | 0.00% | 0.00% | 0.00% | 0.00% | 1.98% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 50% | 2.62% | -0.12% | 100.00% | 0.00% | 0.00% | 0.00% | 3.97% | 0.00% | 4.55% | 0.00% | 0.00% | 0.00% | 7.14% |
| 75% | -2.56% | 0.00% | 0.00% | 0.00% | 0.00% | 100.00% | -2.56% | 0.00% | -3.00% | 0.00% | 100.00% | 0.00% | 0.89% |
| max | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

The next important concern is how easy and feasible it is to obtain the attribute.

|  |  |
| --- | --- |
| Attribute | Test/way to obtain |
| blood urea | BLOOD UREA |
| age | Doctor's Inspection |
| blood pressure | Doctor's Inspection |
| hypertension | Doctor's Inspection |
| coronary artery disease | Doctor's Inspection |
| appetite | Doctor's Inspection |
| pedal edema | Doctor's Inspection |
| hemoglobin | FBC |
| white blood cell count | FBC |
| red blood cell count | FBC |
| anemia | FBC |
| packed cell volume | FBC |
| diabetes mellitus | FBS |
| blood glucose random | RBS |
| serum creatinine | SERUM CREATININE |
| sodium | SERUM ELECTROIDES |
| potassium | SERUM ELECTROIDES |
| pus cell clumps | UFR |
| specific gravity | UFR |
| albumin | UFR |
| sugar | UFR |
| red blood cells | UFR |
| pus cell | UFR |
| bacteria | UFR |

When consider Parameters like “appetite”,



The Distribution of the parameter with the class is bias towards “good appetite” (CKD is not the only reason to have a poor appetite), which will mislead the predictions when apply the training model to a new scenario.

Therefore, by considering all 3 factors mentioned above the "hemoglobin", "specific gravity", "albumin", "hypertension", "diabetes mellitus", "blood glucose random" and “serum creatinine” has been selected as the optimal subset of parameters to do the prediction.

Training the system,

Initially 11 different classification models were training a Logistic Regression, KNN Regression, SVC with a Linear kernel, SVC with RBF kernel, Gaussian NB, a Decision Tree Classifier, Random Forest Classifier, XGB Classifier, Extra Trees Classifier, an Ada Boost Classifier and a Classical Neural Network. Then the dataset has been divided into 3 parts as 70% Training data, 15% cross validation data and 15% test data randomly.

Thereafter, the models were optimized for the Training dataset and has done the initial accuracy and error test from the cross-validation dataset and tested. From the above 11 algorithms, 6 algorithms out perform in Training accuracy, Testing accuracy and in cross validation accuracy. Which are the Decision Tree Classifier, Random Forest Classifier, XGB Classifier, Extra Trees Classifier, Ada Boost Classifier and Classical Neural Network.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Training Accuracy | CV Accuracy | Testing Accuracy |
| Logistic Regression | 96.07% | 96.66% | 95.00% |
| KNN | 97.85% | 98.33% | 98.33% |
| SVC Linear | 97.14% | 96.66% | 96.66% |
| SVC RBF | 94.64% | 95.00% | 95.00% |
| Gaussian NB | 95.35% | 95.00% | 93.33% |
| Decision Tree Classifier | 100% | 100% | 100% |
| Random Forest Classifier | 100% | 100% | 100% |
| XGB Classifier | 99.28% | 100% | 100% |
| Extra Trees Classifier | 100% | 100% | 100% |
| Ada Boost Classifier | 100% | 100% | 100% |
| Classical Neural Network | 97.81% | 97.50% | 97.50% |

|  |  |  |  |
| --- | --- | --- | --- |
| *Algorithm* | *Precision* | *Recall* | *F1-Score* |
| *Logistic Regression* | 1.000 | 0.925 | 0.961 |
| *KNN* | 1.000 | 0.975 | 0.987 |
| *SVC Linear* | 1.000 | 0.950 | 0.974 |
| *SVC RBF* | 1.000 | 0.925 | 0.961 |
| *Gaussian NB* | 0.973 | 0.925 | 0.948 |
| *Decision Tree Classifier* | 1.000 | 1.000 | 1.000 |
| *Random Forest Classifier* | 1.000 | 1.000 | 1.000 |
| *XGB Classifier* | 1.000 | 1.000 | 1.000 |
| *Extra Trees Classifier* | 1.000 | 1.000 | 1.000 |
| *Ada Boost Classifier* | 1.000 | 1.000 | 1.000 |
| *Classical Neural Network* | 0.962 | 1.000 | 0.981 |

Based on the above results the, the algorithms which gave perfect accuracy in all 3 data sets were selected.

Moreover, even though models gave 100% its important to check which attributed make the highest impact on each model to take the decision.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature Importance | | | | | |
| Attribute | Decision Tree Classifier | Random Forest Classifier | XGB Classifier | Extra Trees Classifier | Ada Boost Classifier |
| Hemoglobin | 0.580 | 0.246 | 0.252 | 0.174 | 0.330 |
| Specific Gravity | 0.265 | 0.275 | 0.135 | 0.242 | 0.320 |
| Serum Creatinine | 0.031 | 0.160 | 0.500 | 0.057 | 0.000 |
| Albumin | 0.103 | 0.196 | 0.089 | 0.158 | 0.140 |
| Hypertension | 0.000 | 0.051 | 0.000 | 0.192 | 0.130 |
| Diabetes Mellitus | 0.000 | 0.026 | 0.000 | 0.130 | 0.080 |
| Blood Glucose Random | 0.022 | 0.046 | 0.024 | 0.048 | 0.000 |

After identifying the importance of selected features for each prediction algorithm the standard deviation of the feature importance of each algorithm has calculated, which explicit the algorithm’s bias towards different features.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | Decision Tree Classifier | Random Forest Classifier | XGB Classifier | Extra Trees Classifier | Ada Boost Classifier |
| Standard Deviation | 0.214295746 | 0.102219419 | 0.181369263 | 0.070684746 | 0.136224604 |

When consider the standard deviations and above graph its clear that Extra Trees Classifier has the lowest bias towards features and next the Random Forest Classifier. Decision Tree Classifier has the highest bias of all.