

**Prediction of Chronic Kidney Disease**

by

**Anitha D Praveen**

Data Science Investigation

January, 2020

## **Introduction**

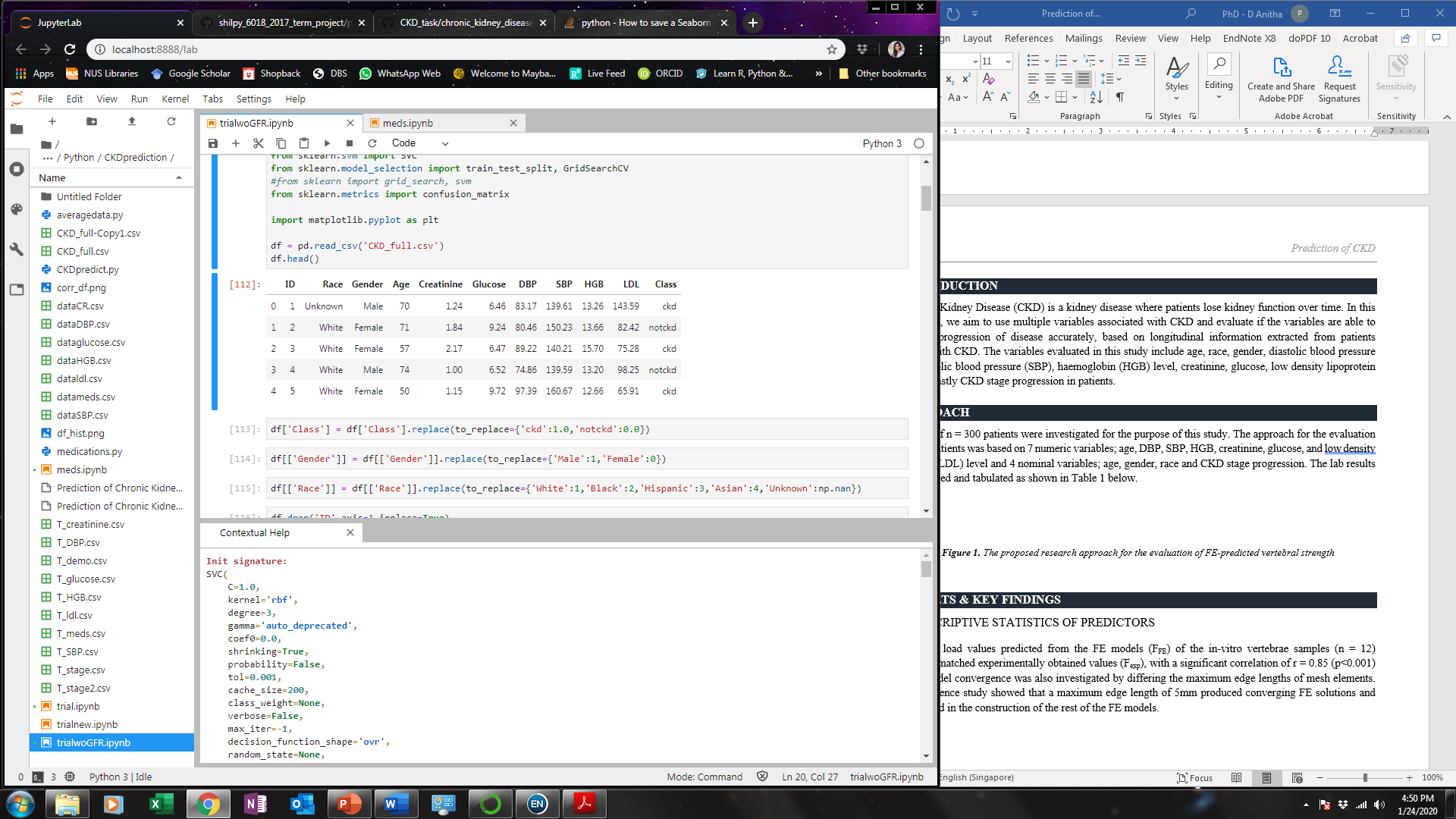
Chronic Kidney Disease (CKD) is a kidney disease where patients lose kidney function over time. In this investigation, we aim to use multiple variables associated with CKD and evaluate if the variables are able to predict the progression of disease accurately, based on longitudinal information extracted from patients diagnosed with CKD. The variables evaluated in this study include age, race, gender, diastolic blood pressure (DBP), systolic blood pressure (SBP), haemoglobin (HGB) level, creatinine, glucose, low density lipoprotein (LDL) and lastly CKD stage progression in patients.

Estimated glomerular filtration (eGFR) is the current gold standard in diagnosing CKD as it is considered the best measure of kidney function. It is primarily determined by the serum creatinine and the preferred equation for the estimation of eGFR is the body surface area-normalized, 4-variable, Modification of Diet in Renal Disease Study (MDRD) Equation based on SCr, age, gender, and ethnicity. However, basing CKD diagnosis just on eGFR alone has not been effective as it has low specificity and sensitivity in detecting patients with this kidney disease. Hence there is a need to integrate more clinical risk factors to better predict the disease.

## **APPROACH**

A total of n = 300 patients were investigated for the purpose of this study. The approach for the evaluation of CKD in patients was based on 7 numeric variables; age, DBP, SBP, HGB, creatinine, glucose, and low density lipoprotein (LDL) level and 4 nominal variables; age, gender, race and CKD stage progression, or also referred to as ‘Class’ in this study. ‘Ckd’ refers to patients who progressed to the next stage of CKD while ‘notckd’ refers to patients who did not progress to the next stage. The lab results were processed and tabulated as shown in Table 1 below.

***Table 1.*** *The proposed research approach for the evaluation of FE-predicted vertebral strength*

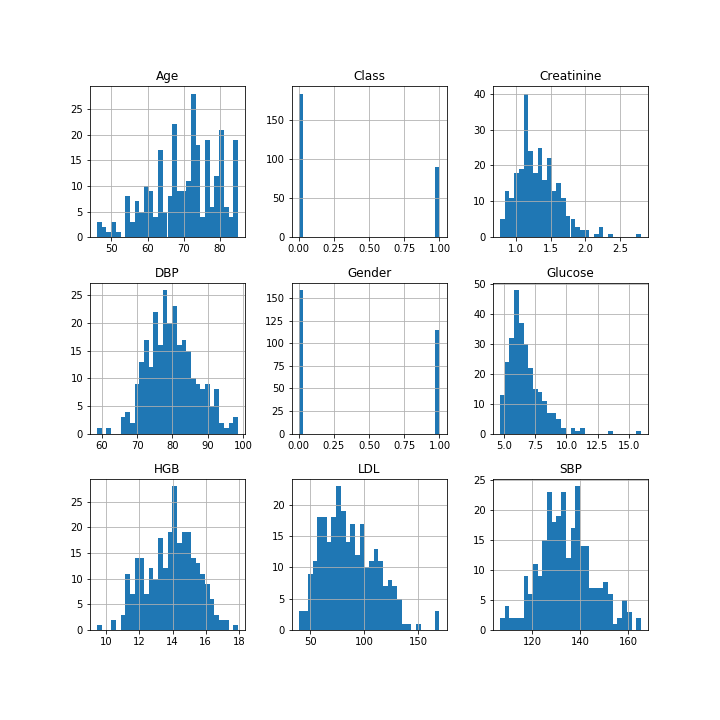


With missing or unknown values dropped from the tabulation, there were a remaining total of 274 patients that were analysed, of which 184 and 90 patients belonged to the ‘notckd’ group and ‘ckd’ group respectively. These variables were correlated and compared to observe the inter-correlations and then classifier analysis was performed to evaluate the best method to provide the highest prediction accuracy. This method was then used to compare the prediction accuracy of integrating all the clinical risk factors versus using GFR alone to monitor the progression of CKD.

The efficacy of medications were investigated separately, due to the extensive variations involved in drug treatment, in terms of the choice and dosage of medication.

## **Results & key Findings**

## **Descriptive Statistics of predictors**



***Figure 1.*** *Histogram plots of input variables used for CKD disease prediction*

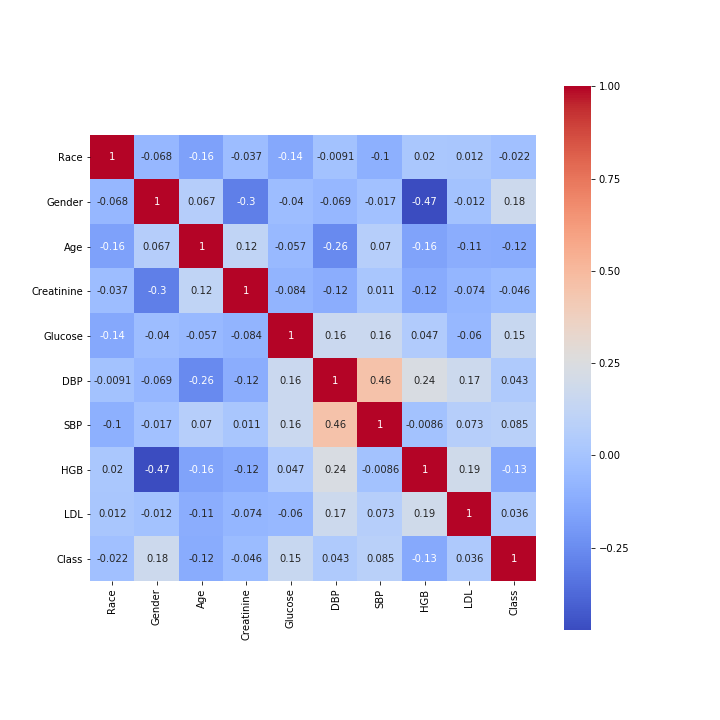
Class, and gender are evidently nominal variables. The following descriptive statistics are reported in mean and standard deviations (SD). The mean age of patients is 70 (±9) years old, with the minimum age of patients at 46 and maximum at 85 years old. DBP. SBP. LDL. HGB, creatinine and glucose exhibit a normal distribution (Fig.1). Mean DBP. SBP. LDL. HGB, creatinine and glucose levels are shown in the table below.

***Table 2.*** *Descriptive statistics of numeric variables used in this study*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Mean (SD)** | **Minimum, Maximum** |
| Age (years) | 70 (9) | 46, 85 |
| DBP (mm/Hg) | 79.7 (6.9) | 58.7, 98.5 |
| SBP (mm/Hg) | 134.0 (11.2) | 106.6, 165.2 |
| LDL (mg/dL) | 86.7 (24.9) | 39.3, 169.9 |
| HGB (gms) | 13.9 (1.5) | 9.4, 18.0 |
| Creatinine (gms) | 1.3 (0.3) | 0.8, 2.8 |
| Glucose (cells/cumm) | 6.7 (1.4) | 4.7, 16.0 |

## **Correlations between predictors**

The Pearson correlations do not exhibit any significant inter-correlations between predictors, with the correlations below 0.5. This demonstrates that the analysed variables are independent of each other.



***Figure 2.*** *Correlations between different variables*

## **Comparison of Classifiers and Prediction accuracies**

## Based on the prediction accuracies, K Neighbors classifier, linear discriminant analysis and logistic regression give the highest accuracies.

## ***Table 1.*** *Classifiers and the respective accuracy predictions and precision scores.*

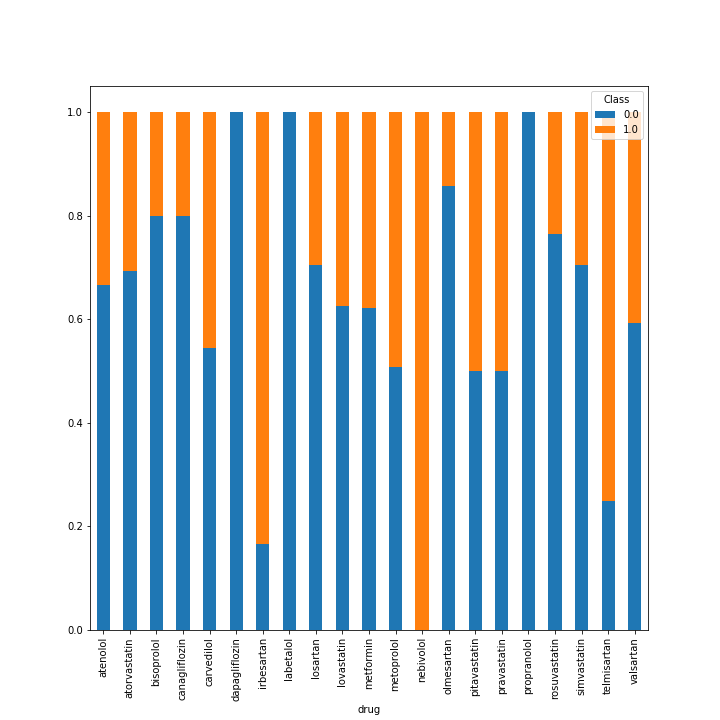
|  |  |  |
| --- | --- | --- |
| **Classifier** | **Accuracy (%)** | **Precision score** |
| K Neighbors Classifier | 65.5 | 0.71 |
| SVC | 60.0 | 0 |
| Nu SVC | 50.9 | 0.39 |
| Decision Tree Classifier | 58.2 | 0.47 |
| Random Forest Classifier | 65.5 | 0..71 |
| AdaBoost Classifier | 61.8 | 0.55 |
| Gradient Boosting Classifier | 70.9 | 0.69 |
| Gaussian NB | 74.5 | 0.79 |
| Linear Discriminant Analysis | 65.5 | 0.67 |
| Logistic Regression | 65.5 | 0.67 |

## **Effect of Medication on CKD Progression**

Amongst this group of patients analysed for the purpose of this study, n = 21 different drugs were administered. Atorvastatin and metformin is the most commonly administered medications, with 44% of atorvastatin patients and an alarming 61% of metformin patients progressing to the next stage of CKD. Similar results are also shown qualitatively in Figure 3.

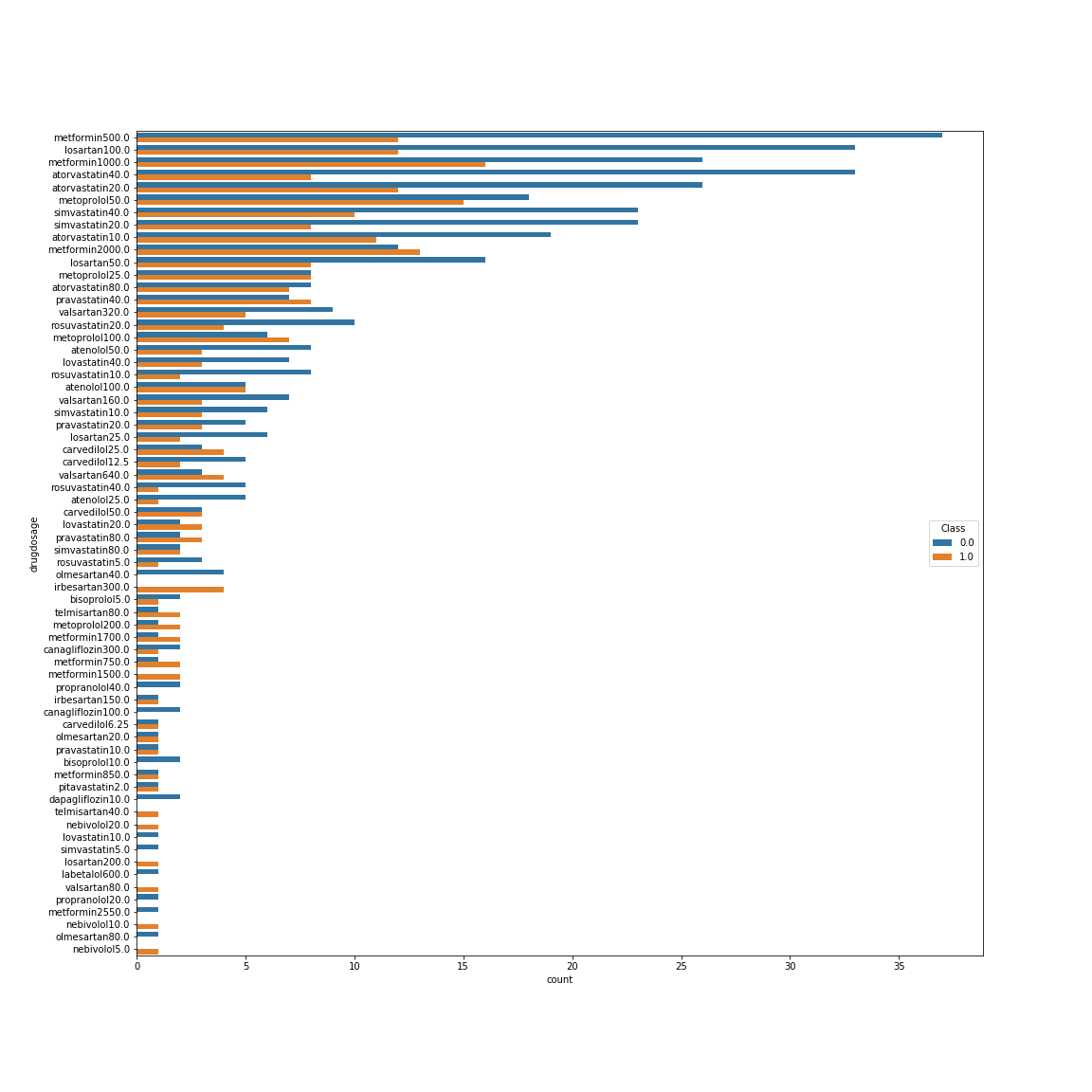
***Table 3.*** *Different medications administered to patients and their respective CKD stage progression with Class = 0 representing patients who did not progress to next stage and Class = 1 representing patients who did progress to next stage*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Drug** | **Class** | **No. of Patients** | **Drug** | **Class** | **No. of Patients** |
| **Atenolol** | 0 | 18 | **Metoprolol** | 0 | 33 |
| 1 | 9 | 1 | 32 |
| **Atorvastatin** | 0 | 86 | **Nebivolol** | 1 | 3 |
| 1 | 38 | **Olmesartan** | 0 | 6 |
| **Bisoprolol** | 0 | 4 | 1 | 1 |
| 1 | 1 | **Pitavastatin** | 0 | 1 |
| **Canagliflozin** | 0 | 4 | 1 | 1 |
| 1 | 1 | **Pravastatin** | 0 | 15 |
| **Carvedilol** | 0 | 12 | 1 | 15 |
| 1 | 10 | **Propranolol** | 0 | 3 |
| **Dapagliflozin** | 0 | 2 | **Rosuvastatin** | 0 | 26 |
| **Irbesartan** | 0 | 1 | 1 | 8 |
| 1 | 5 | **Simvastatin** | 0 | 55 |
| **Labetalol** | 0 | 1 | 1 | 23 |
| **Losartan** | 0 | 55 | **Telmisartan** | 0 | 1 |
| 1 | 23 | 1 | 3 |
| **Lovastatin** | 0 | 10 | **Valsartan** | 0 | 19 |
| 1 | 6 | 1 | 13 |
| **Metformin** | 0 | 79 |  |  |  |
| 1 | 48 |  |  |  |



***Figure 3.*** *Bar plot of the various drugs administered to patients with Class = 0 (blue) representing patients who did not progress to next stage and Class = 1 (orange) representing patients who did progress to next stage*

Looking at the medications and dosage closer, metformin with a daily drug dosage of 500 was the most commonly administered to patients. Due to the differences in number of drugs that were administered to this group of patients, confounded by the fact that there were various daily dosages, it is not possible to statistically determine the efficacy of some drugs as the subgroup of patients receiving these drugs were too small.



***Figure 4.*** *Bar plot of the various drugs administered to patients with Class = 0 (blue) representing patients who did not progress to next stage and Class = 1 (orange) representing patients who did progress to next stage*