**CHRONIC KIDNEY (RENAL) DISEASE PREDICTION USING MACHINE LEARNING**

Selva Vignesh M,

MSc Data Science,

Department of Data Science,

Coimbatore Institute of Technology.

saiselva2302@gmail.com

**Abstract:**

In today's era everyone is trying to be conscious about health even through the workload and busy schedule, one gives attention to the health when it shows any symptoms of some kind. In order to that, the field of biosciences have advanced to a larger extent and have generated large amounts of information from Electronic Health Records. This have given rise to the acute need of knowledge generation from this enormous amount of data. Data mining methods and machine learning play a major role in this aspect of biosciences. Chronic Kidney Disease (CKD) is a condition in which the kidneys are damaged and cannot filter blood as they always do. A family history of kidney diseases or failure, high blood pressure, type 2 diabetes may lead to CKD. This is a lasting damage to the kidney and chances of getting worse by time is high. The very common complications that results due to a kidney failure are heart diseases, anaemia, bone diseases, high potassium and calcium. The worst case situation leads to complete kidney failure and necessitates kidney transplant to live. An early detection of CKD can improve the quality of life to a greater extent. This calls for good prediction algorithms to predict CKD at an earlier stage. A wide range of machine learning algorithms are employed for the prediction of CKD. The techniques which will be used here are data pre-processing, exploratory analysis of data, model optimization, various classifiers to predict CKD, and finally the best classifier will be chosen, which can predict CKD in a very early stage.

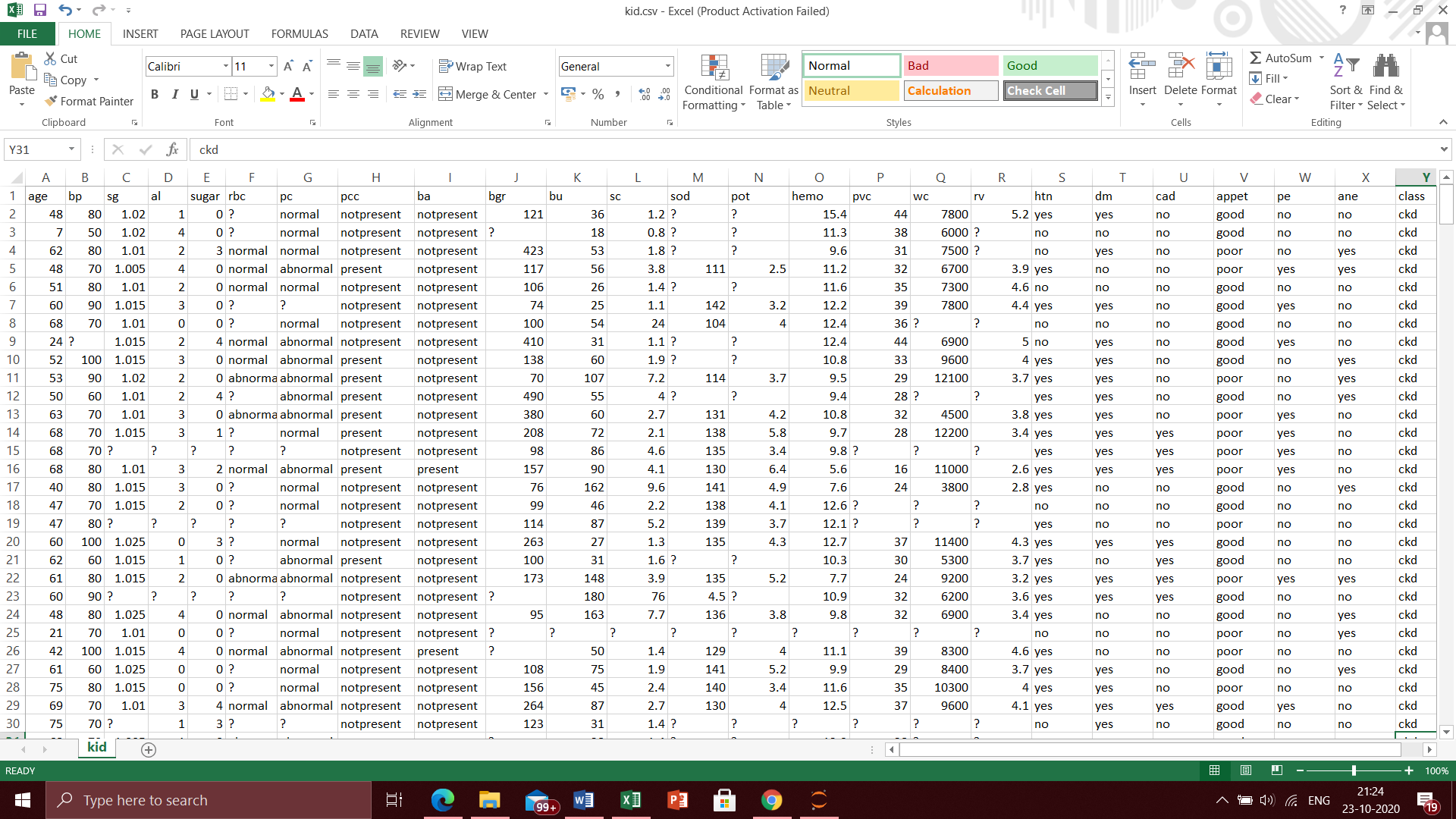
***Keywords:*** CKD, Machine Learning, Algorithms

**Introduction:**

We all knew that Kidney is essential organ in human body. Which has main functionalities like excretion and osmoregulation. In simple words we can say that all the toxic and unnecessary material from the body is collected and thrown out by kidney and excretion system. There are approximately 1 million cases of Chronic Kidney Disease (CKD) per year in India. Chronic kidney disease is also called renal failure. It is a dangerous disease of the kidney which produces gradual loss in kidney functionality. CKD is a slow and periodical loss of kidney function over a period of several years. A person will develop permanent kidney failure. If CKD is not detected and cured in early stage then patient can show following Symptoms: Blood Pressure, anaemia, poor nutrition health and nerve damage, Decreased immune response because at advanced stages dangerous levels of fluids, electrolytes, and wastes can build up in your blood and body. Hence it is essential to detect CKD at its early stage but it is unpredictable as its Symptoms develop slowly and aren't specific to the disease. Some people have no symptoms at all so machine learning can be helpful in this problem to predict that the patient has CKD or not. Machine learning does it by using old CKD patient data to train predicting model. Glomerular Filtration Rate (GFR) is the best test to measure your level of kidney function and determine your stage of chronic kidney disease. It can be calculated from the results of your blood creatinine, age, race, gender, and other factors. The earlier disease is detected the better chance of showing or stopping its progression.

**Sample Dataset:**

**Source:** https://archive.ics.uci.edu/ml/datasets/chronic\_kidney\_disease



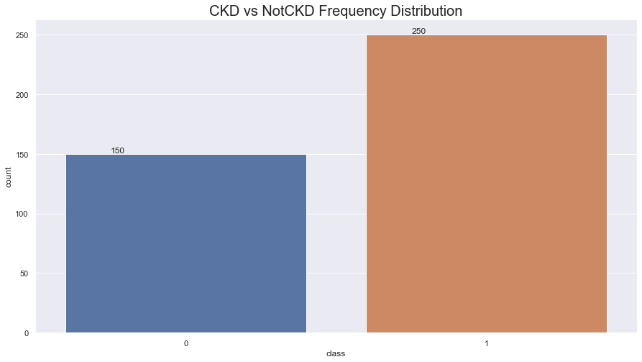
**Data Pre-processing:**

First we check for null values in the dataset. Here our dataset is yet to be cleaned, so we convert the “?” symbols and null values to Nan. Then we analyse each and every column of the dataset. The first column age consists of Nan values, which we replace it with the mean of that particular column. Similarly it is done for the blood pressure, specific gravity, albumin, Blood glucose random, Blood urea, Serum creatinine, Sodium, Potassium, Haemoglobin, Packed cell volume, White blood cells count and Red blood cell count. The sugar column consists of values from 0 to 5, where there are missing values too. As finding mean is an inappropriate method, we use the mode method to replace the Nan values. The mode method replaces the Nan values with the values that are repeated number of times. It is repeated for variables red blood corpuscles, pus cells, pus cell counts, bacteria, hypertension, diabetes, coronary artery disease, appetite, pedal edema and anaemia. The Red blood corpuscle is a nominal column which consists of normal and abnormal results. We convert the categorical variables to numerical variables (normal – 1, abnormal – 0). We convert the nominal variables to have maximum number of 1s to reduce the zero values too. Similarly we convert the pus cell and puss cell count attributes too. Likewise, the bacteria variable is replaced to 0s and 1s from present and non-present. The hypertension variable consisting of yes and no is replaced with 0s and 1s. The same is done for anaemia too. The appetite variable consisting of no, poor and good is replaced with 0, 1 and 2. Similarly for the pedal edema variable too. Now our data pre-processing is over and now the cleaned data is ready for data analysis.

**Exploratory Data Analysis:**

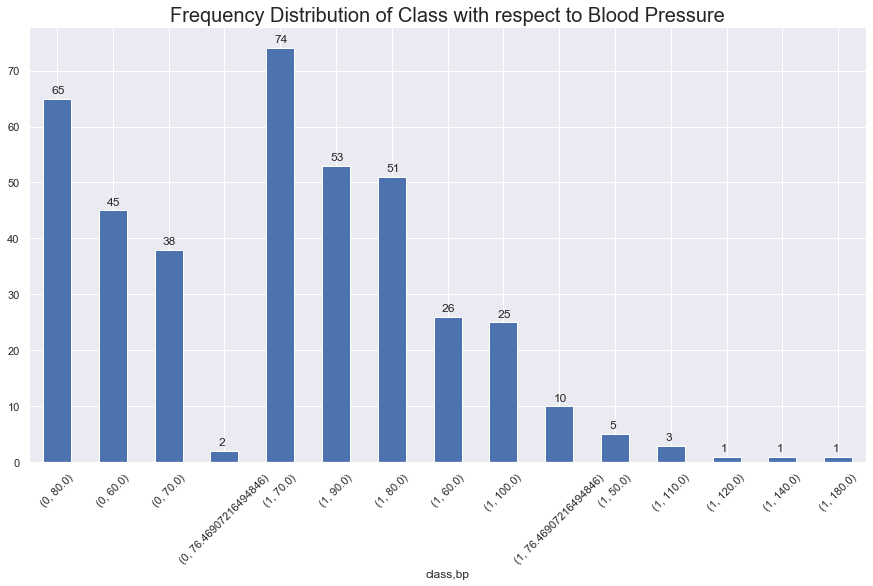
Exploratory Data Analysis (EDA) is an approach for analysing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modelling task. For better understanding of our chronic dataset well, we will also be performing EDAs in such a way that we can clearly see the dependencies and weightage of values clearly. After understanding the variables, it will be easy for us to implement them in the model.

**Class variable:**

****

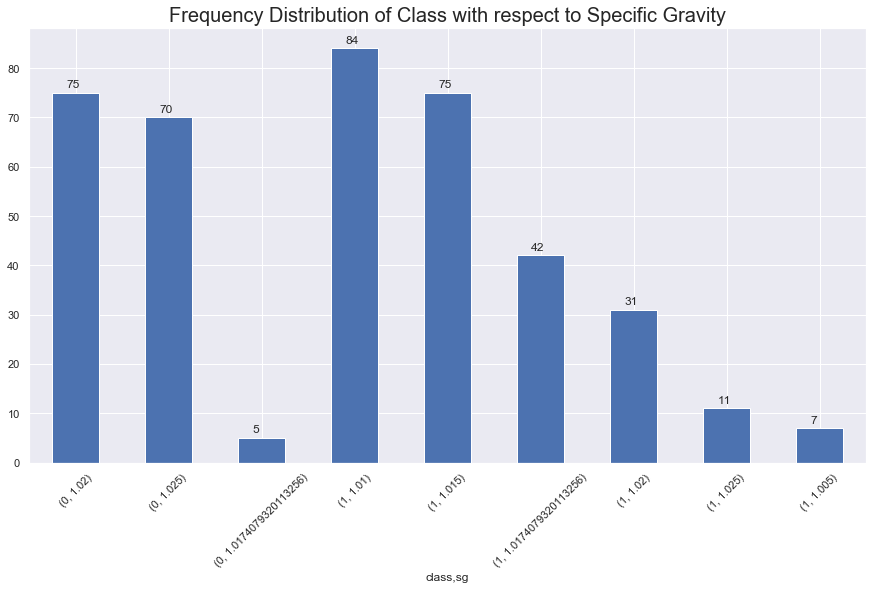
From the bar plot we can see that there are many ckd positive patients (250) than the ckd negative patients (150). The dataset is dominated by the positive patients, so we will develop a model that clearly predicts the outcome.

**Blood pressure VS Class:**



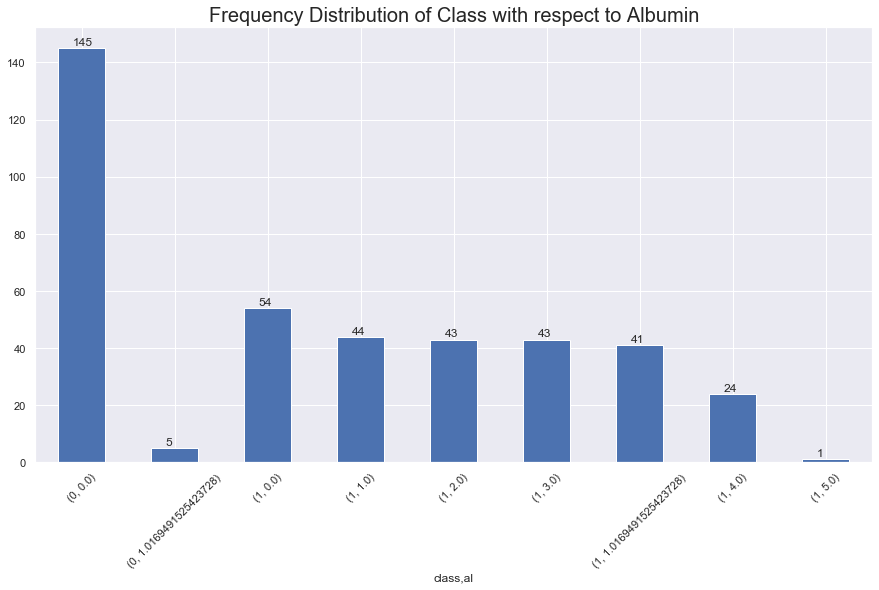
Here the plot clearly show that many patients fall on the range of 70-80 for min BP and very few patients on the scale of 100-120 with mac BP. Having a high BP is a risk factor for CKD.

**Specific gravity VS Class:**



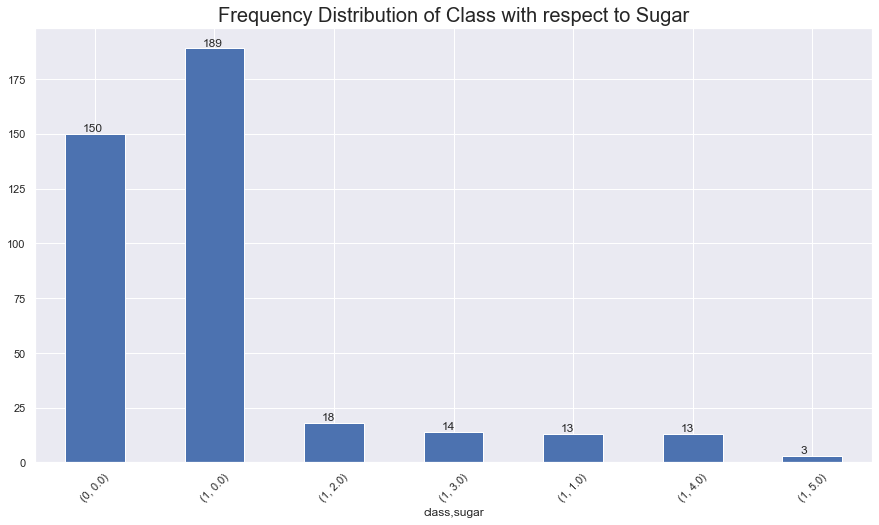
The specific gravity is nothing but the density of the urine. It normally ranges from 1.002-1.030 sg. Here we can see that many people have density ranging from 1.01-1.02. Only a few patients fall above the range. High specific gravity leads to kidney stones, so they have to be detected early and given treatment.

**Albumin VS Class:**



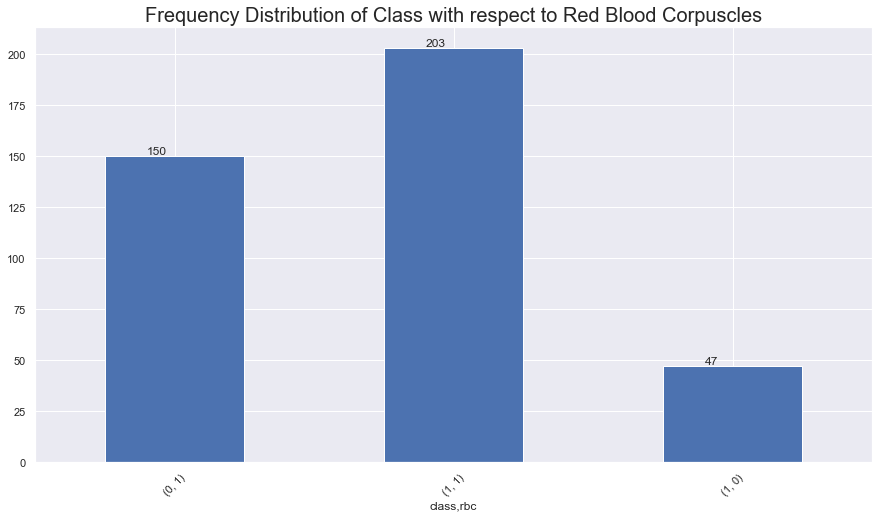
Albumin is the mixture of fatty acid, bilirubin, ions etc. High content of albumin in urine causes renal failure. Here we can see that there are many patients in the lower limit. So there are only few patients who have to be treated in order to reduce the albumin content in urine.

**Sugar VS Class:**

****

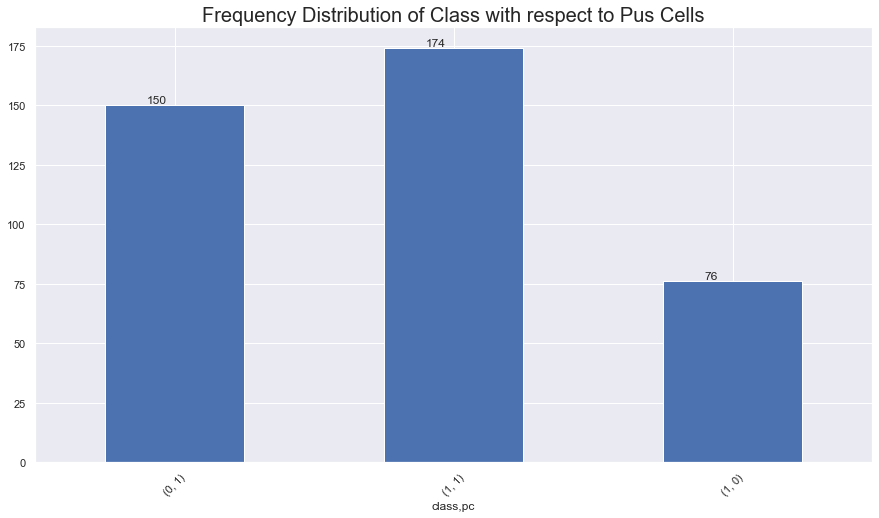
The Sugar has a range of 0-5 in out dataset, where most of the patients have a high sugar as the fall on 0-1. Only a few patients have mild sugar content in their blood. So there are chances for diabetes among the patients.

**Red blood corpuscles VS Class:**



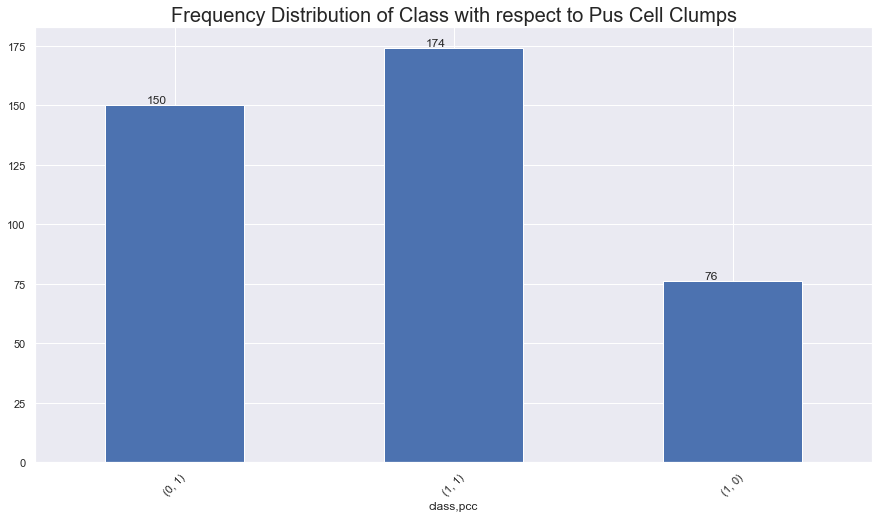
The Red Blood Corpuscles count is one of the main factor in CKD, as it is essential for filtering the blood in the kidney. From the plot we can see that half of the patients are normal and half of them are abnormal. So we need to concentrate here to classify the patients effectively, as it is one of the highly contributing variable to the class.

**Pus cells VS Class:**



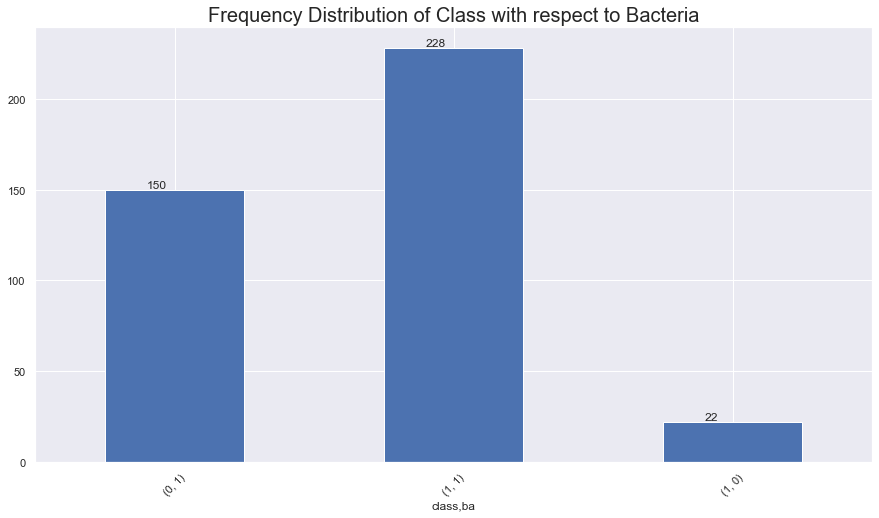
Pus cells causes urinary tract infection which makes a cloudy appearance in the urine. This is most likely to affect women than men. Here we can see that three fourth of the patients have puss cells and only we have the absence of it. So they have to be treated to remove the puss cells because they can cause severe urinary tract infections.

**Pus cell clumps VS Class:**

****

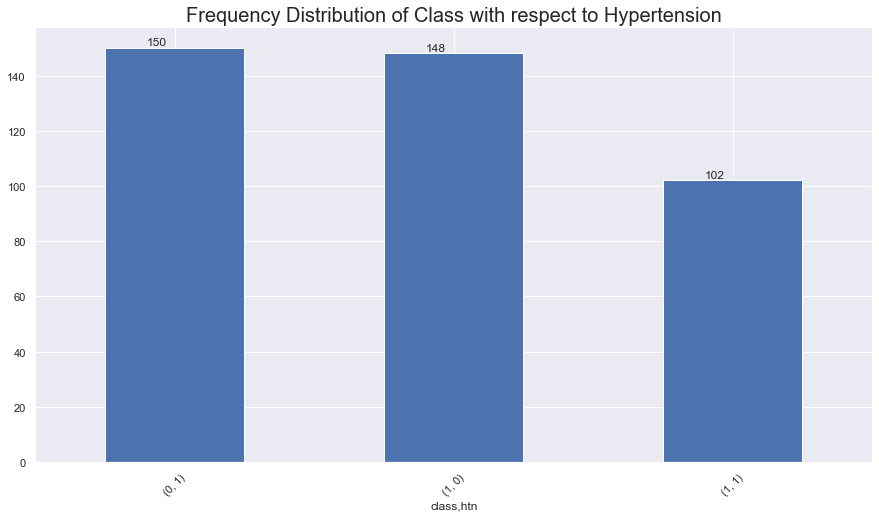
The puss cell clumps are nothing but the inflammation caused in the urinary tract due to puss cells. They are highly dangerous because they could block the urinary tract. This is same to that of the pus cells as they both are related. Here also ¾ of the patients have puss cell clumps.

**Bacteria VS Class:**



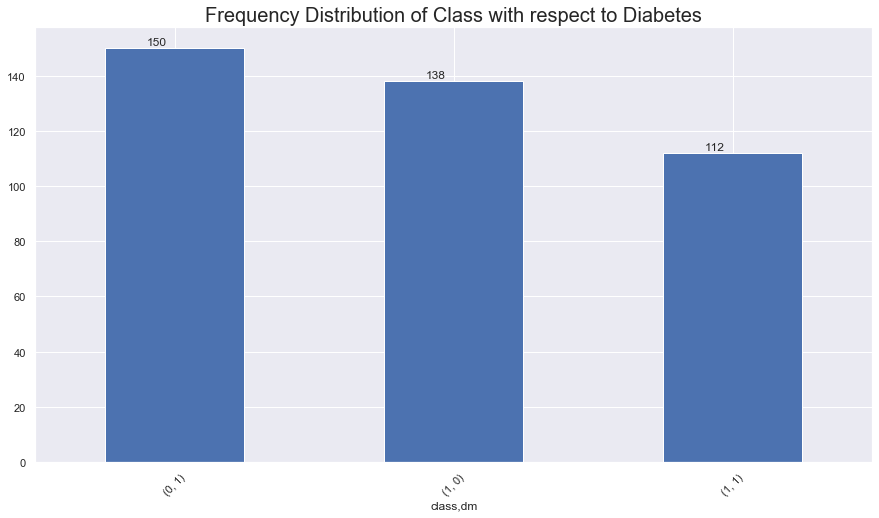
Bacterial presence in the urine causes allergy and irritation in the urinary track. It also blocks the blood filtration process in the kidney and results in improper flow of urine. Here bacterial effect seems to be low on maximum of the patients, but they are affected by other factors which are also the types of bacteria in the urea.

**Hypertension VS Class:**

****

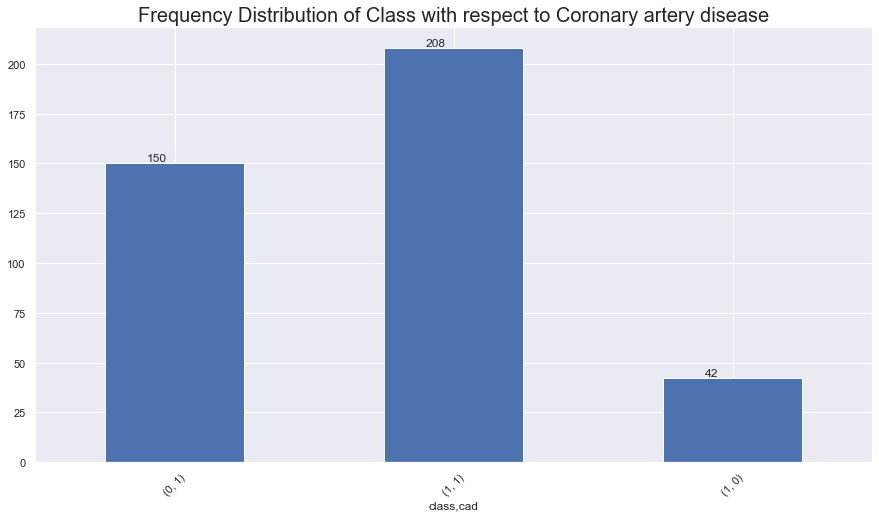
Hypertension is the second leading cause of kidney failure. Over time, uncontrolled high blood pressure can cause arteries around the kidneys to narrow, weaken or harden. These damaged arteries are not able to deliver enough blood to the kidney tissue. As we can see almost majority of the patients have hypertension, which caused them renal failures.

**Diabetes VS Class:**

****

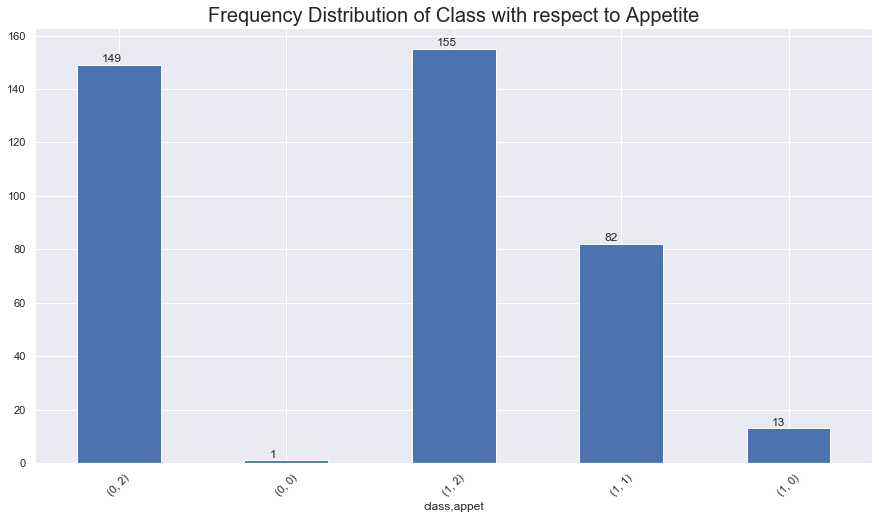
Diabetes is also a risking factor to cause renal problems, as the sugar content in the blood is high. This directly affects the filtration of blood in the kidney. The patients will have continuous flow of urine which causes improper filtration. Here we can see that many patients are affected by diabetes. About 260 of 400 patients have diabetes, so they have to be treated soon.

**Coronary artery disease VS Class:**

****

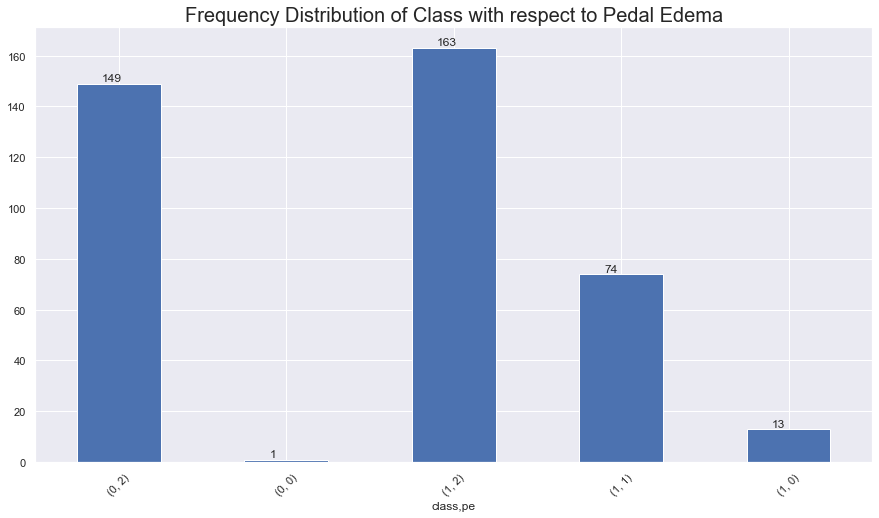
Coronary artery disease is actually an independent risk factor for CKD. When you have heart disease, your heart may not pump blood in the right way. Your heart may become too full of blood. This causes pressure to build in the main vein connected to your kidneys, which may lead to a blockage and a reduced supply of oxygen rich blood to the kidneys. This can lead to chronic kidney disease. Here ½ of the patients have CKD. As it is not a main factor of cause it can be treated in due course of the other attributes itself.

**Appetite VS Class:**

****

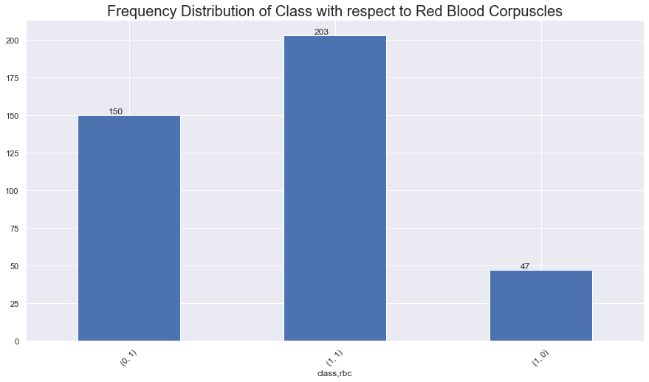
Appetite may worsen with progression of kidney disease leading to malnutrition. As nutrition status is an important factor in dialysis and natural filtration process it is important to have a proper appetite. Here the patients have good, poor and no appetite too. We can see that maximum of the patients have a good appetite with respect to the class variable.

**Pedal Edema VS Class:**



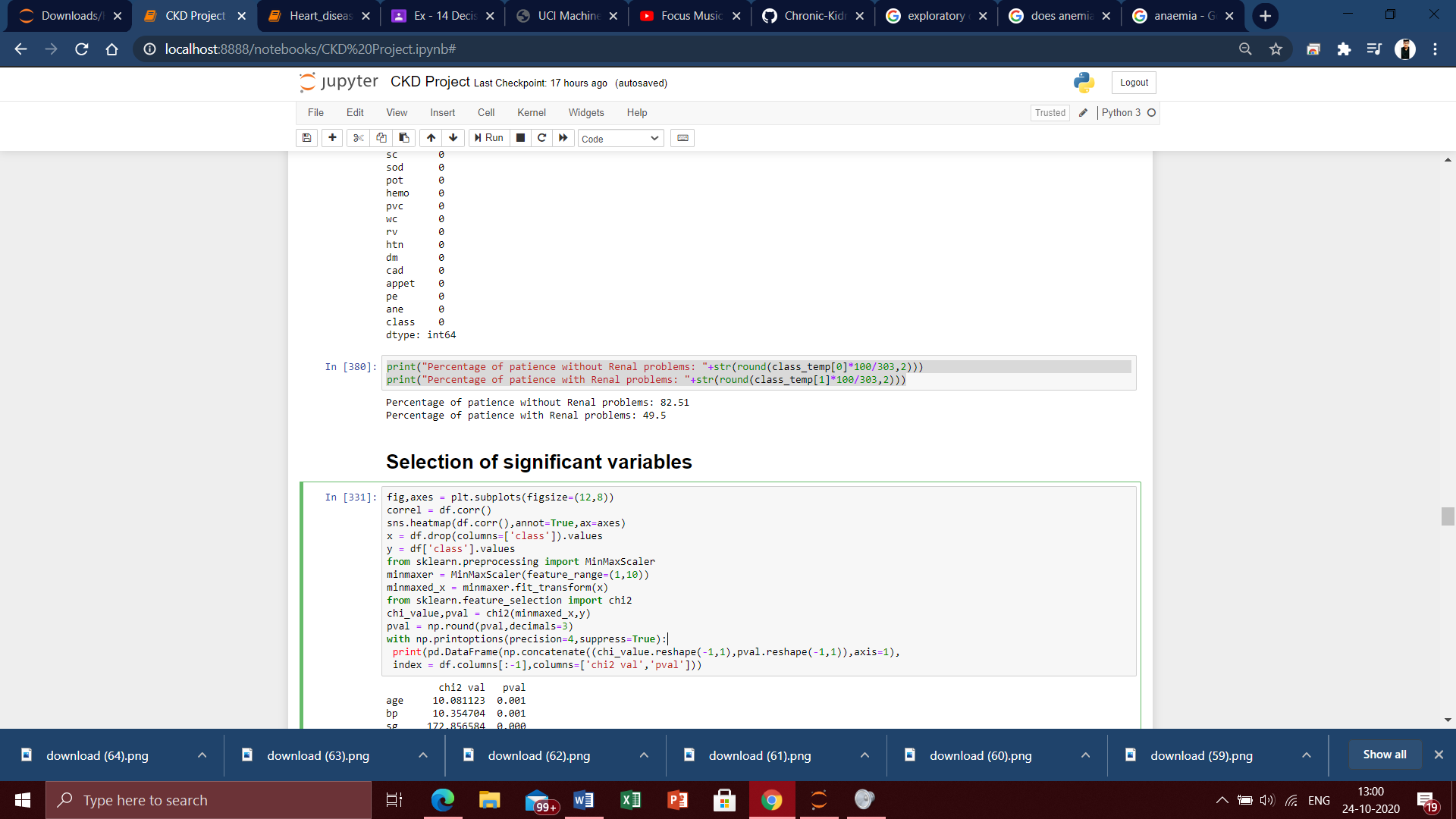
Pedal Edema is quite related to appetite as appetite causes edema around our body. Diuretics are a type of medication that causes the kidneys to excrete more water and sodium, which can reduce edema. Diuretics must be used with care because removing too much fluid too quickly can lower the blood pressure, cause light-headedness or fainting, and impair kidney function. The results of appetite and pedal edema are mostly the same.

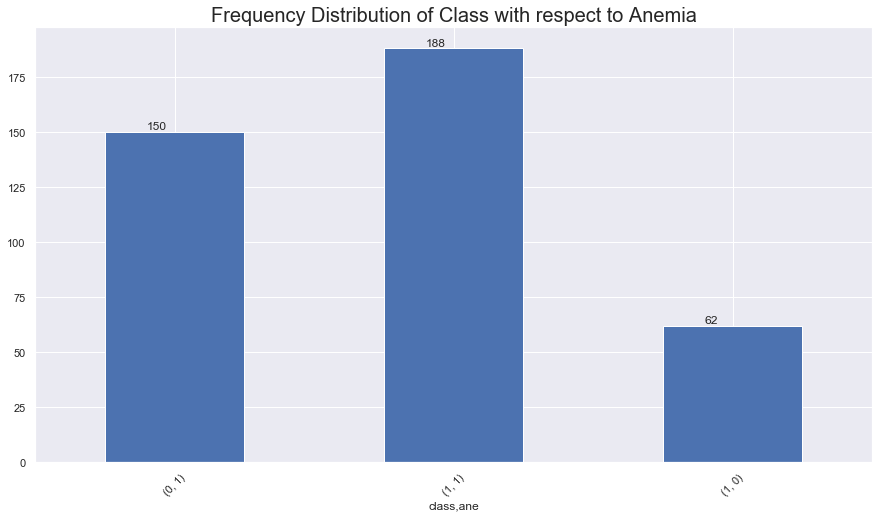
**Red blood corpuscles VS Class:**



The Red Blood Corpuscles count is one of the main factor in CKD, as it is essential for filtering the blood in the kidney. From the plot we can see that half of the patients are normal and half of them are abnormal. So we need to concentrate here to classify the patients effectively, as it is one of the highly contributing variable to the class.

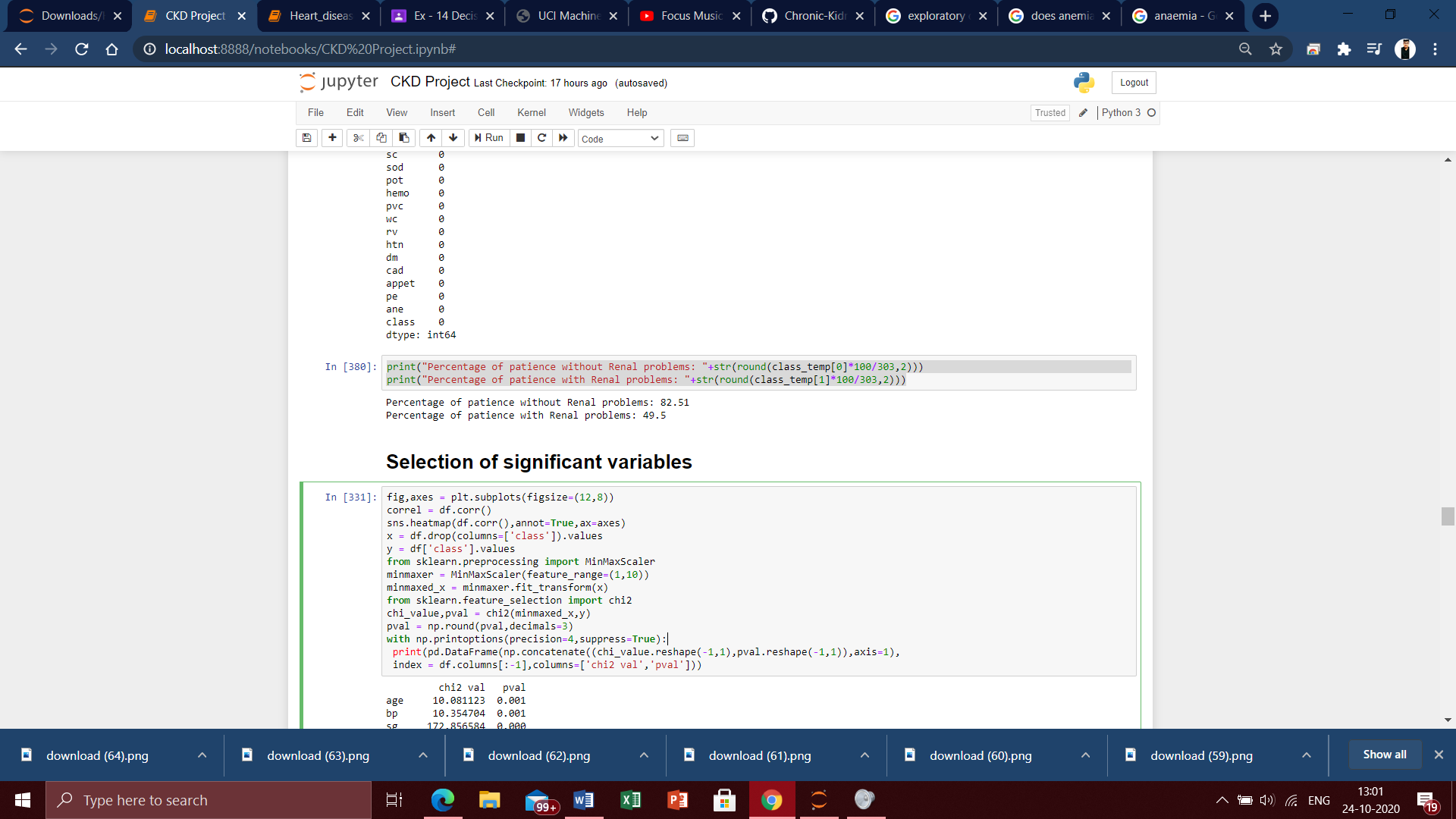
**Anemia VS Class:**



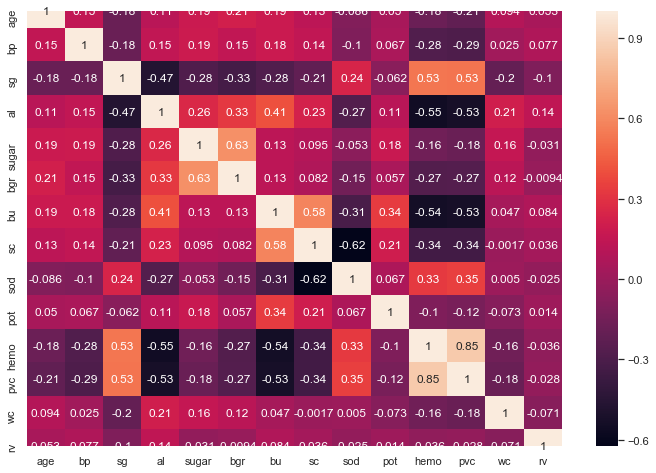


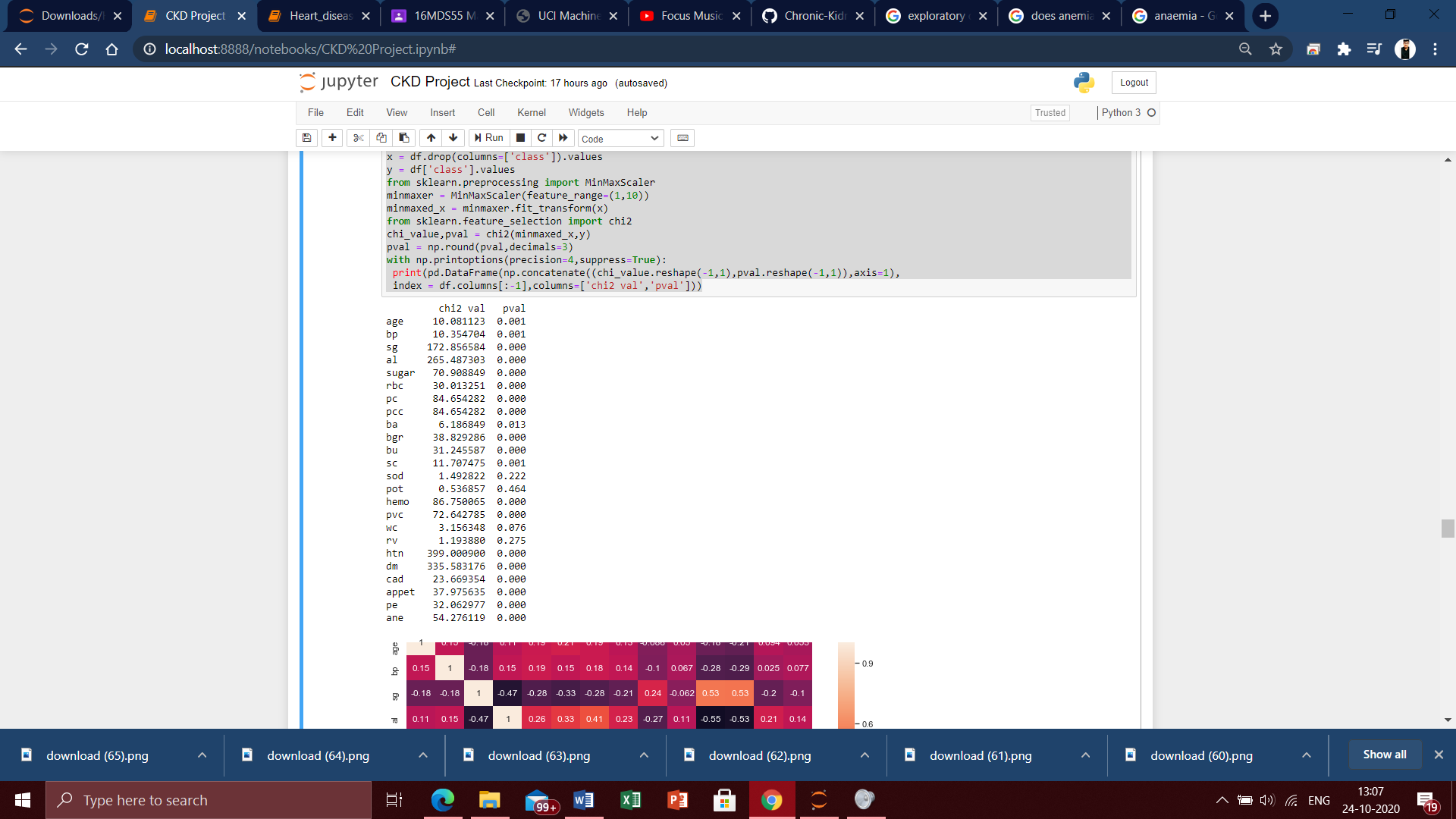
It is a condition in which the blood doesn't have enough healthy red blood cells. When there is improper filtration of blood in our kidney, it causes anemia as most of our blood lacks RBCs and is not purified. This results in the reduced count of RBC which causes anemia. Here we can clearly see that most of the patients are affected by anemia as they had reduced RBC count as we saw before. The patients have to be diagnosed early to avoid CKD.

**Overall inference of patients with and without CKD from dataset:**



**Selection of significant variables:**

****



**Inference:**

Choosing significant variables is an important process before we build a model. Because there will be certain variables which may contain many null and Nan values, they might not contribute much to the model’s accuracy etc. So we will be dropping some variables based on the Chi-squared test. Chi-squared test is used to determine the significant variables. The variables which doesn’t provide significant information to the renal disease classification are removed. P-value can be used to infer how significant the variable contributions are. Variables with p-value greater than 0.10 are removed as they variables doesn’t provide significant contribution to classify the variables at 90% confidence level. The insignificant variables which has been removed for further proceedings are ‘Sodium’, ‘Potassium’, ‘Red blood cell count’. Dropping these variables we will train the new dataset to implement it in the model.

**Splitting dataset into training and test data:**

We have dropped the sodium, potassium and RBC count variables and now we split out processed data into training and test data in the ratio of 70:30.

**Model Optimisation:**

Optimising a model is essential for any kind of dataset as, the model sometimes over fit or under fit to the given data. To reduce those bias, we optimize the model to increase its efficiency and for its true prediction and accuracy.

Here we use the K-fold cross validation method to optimize our model. In K-fold Cross-Validation (CV) we still start off by separating a test set from the remaining data in the data set to use for the final evaluation of our models. The data that is remaining, i.e. everything apart from the test set, is split into **K** number of folds (subsets). The Cross-Validation then iterates through the folds and at each iteration uses one of the K folds as the validation set while using all remaining folds as the training set. This process is repeated until every fold has been used as a validation set. Here we give the CV value as 10 for 10 fold cross validation. By training and testing the model K number of times on different subsets of the same training data we get a more accurate representation of how well our model might perform on data it has not seen before. In a K-fold CV we score the model after every iteration and compute the average of all scores to get a better representation of how the model performs compared to only using one training and validation set.

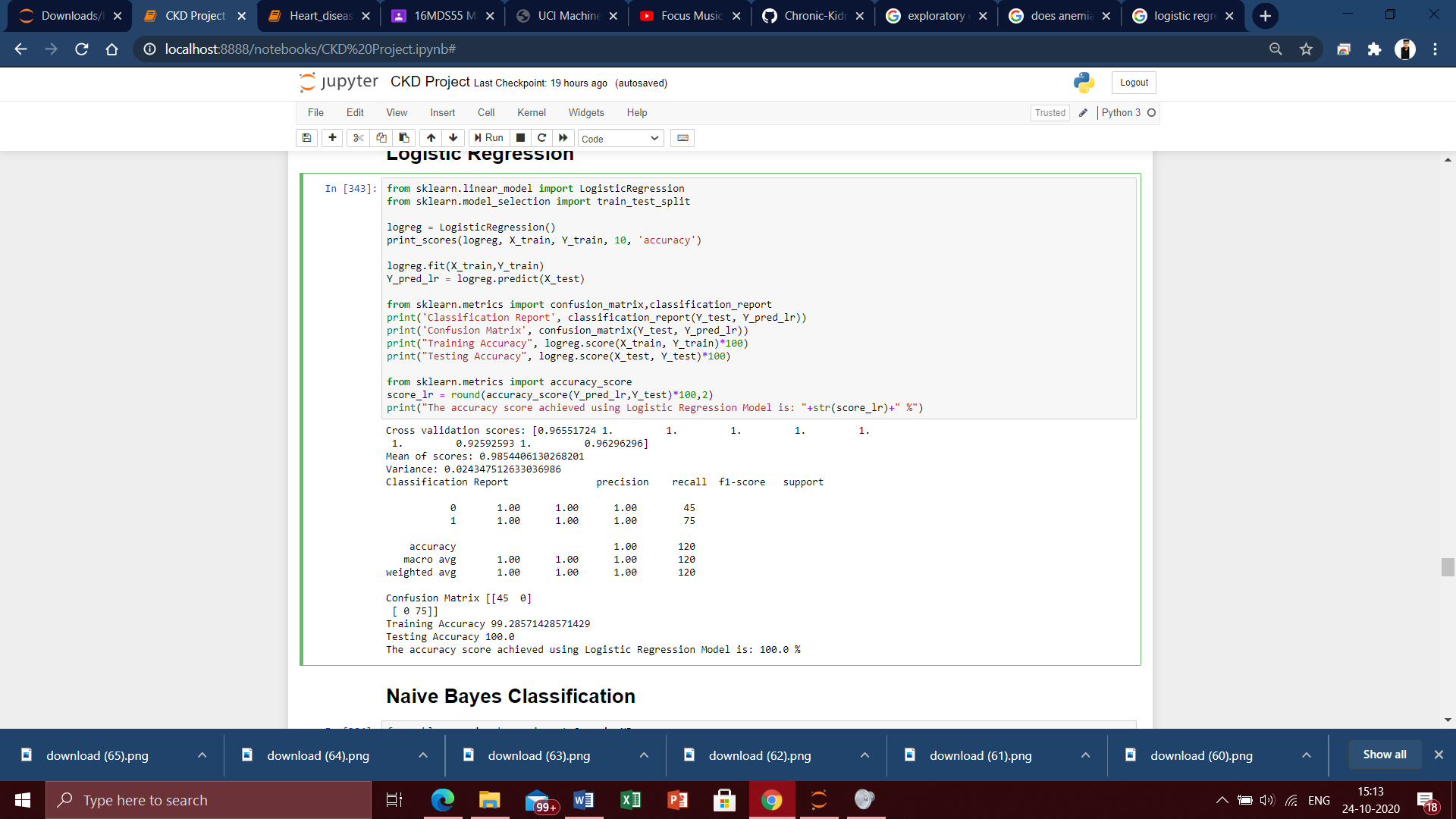
**Fixing the models:**

Now we are into the main part of our project i.e. model fixing. We will be training our models in order to make the best prediction. The Chronic kidney disease dataset is a classification dataset, so we will be using 6 types of classifiers here and at last we will be finding the best models for classification. The classifiers used here are,

* Logistic Regression
* Naive Bayes classification
* Support Vector Machines
* KNN Classifier
* Decision Tree
* Random Forest

**Logistic Regression:**

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. It is a widely used model to analyse the relationship between multiple independent variables and one categorical dependent variable. Here the first used classifier for the CKD prediction is Logistic regression, which is also a type of regression. After fixing the model we can see how it performs by looking into its accuracy.



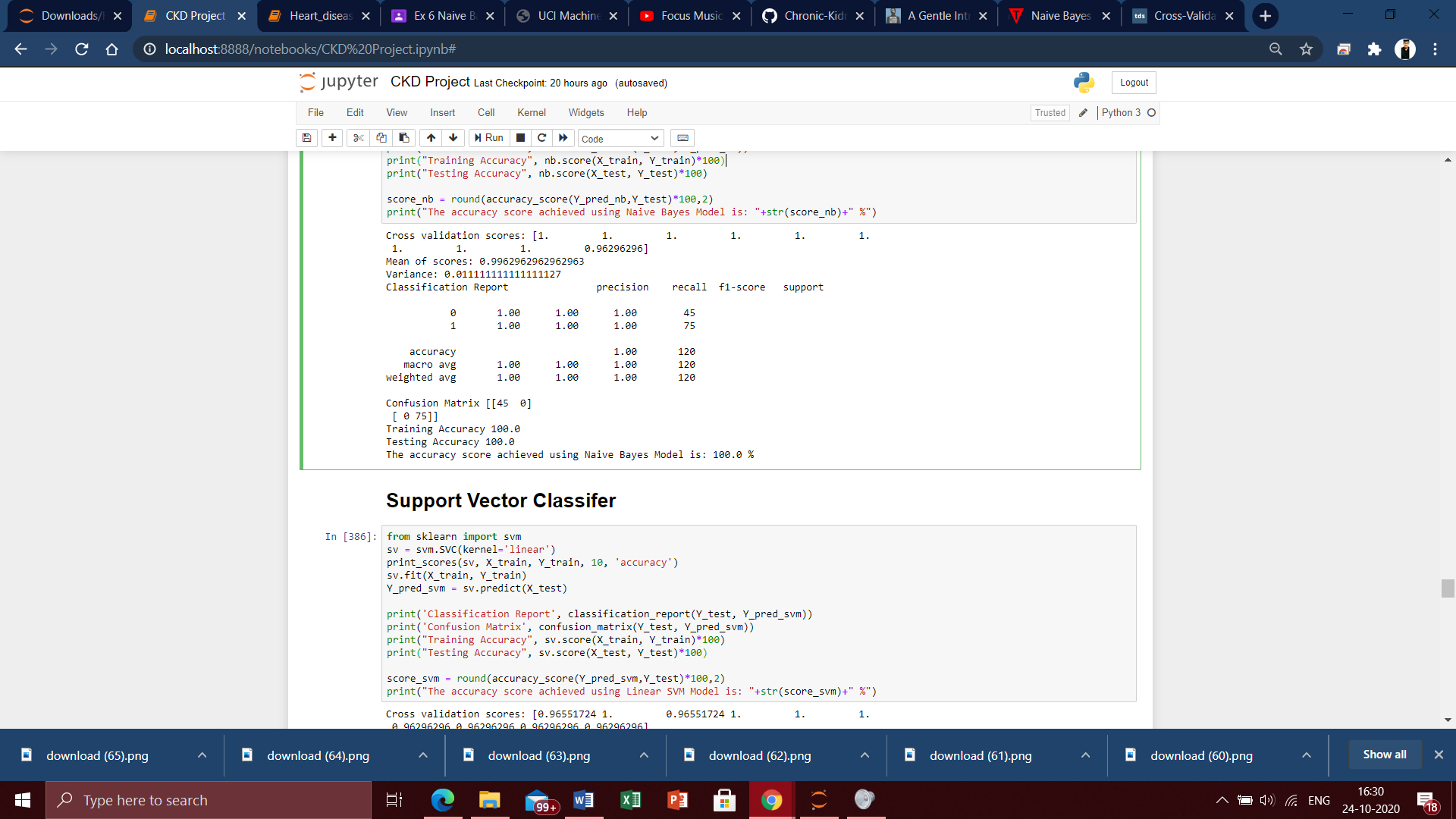
From the logistic regression model we have first obtained the K fold Cross validation scores for about 10 folds. Majority of the scores are 1 and there are chances that our model’s accuracy will be 100%. Next the mean of the scores we obtained that is about 0.98544, variance of about 0.02434. Here the variance is so low so there are no errors occurred in our model. Now comes the classification report which displays the precision, recall, F1 and support scores for the model.

* Precision Score for class 0 (negative) is 1.00 and for class 1 (positive) is 1.00, indicating the preciseness of the model which is so accurate.
* Recall value for class 0 is 1.00 and class 1 is 1.00, which describes the amount up – to which the model can predict the output.
* As the precision and recall values are similar, there are no dominance in classes.

Confusion matrix is nothing but an error matrix. From the matrix we can see that there are 45 true positives, 0 false positive and false negative and finally 75 true negative values. The model didn’t even make a slight error and predicted the values perfectly. The training and test accuracy of the model are 99.25% and 100% respectively. Finally we have come to the accuracy of the model. Surprisingly, we have the model accuracy of 100%. The model has worked well and classified the patients with and without CKD with an accuracy of 100%. The model is also not over-fitted as model optimisation techniques like K fold cross validation is done for 10 folds. So the Logistic regression model can effectively classify the patients without even a slightest error.

**Naive Bayes Classification:**

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. Here for our chronic kidney disease dataset, we will be using the Naive Bayes classification to see how this model works on classifying the patients with CKD and not CKD.



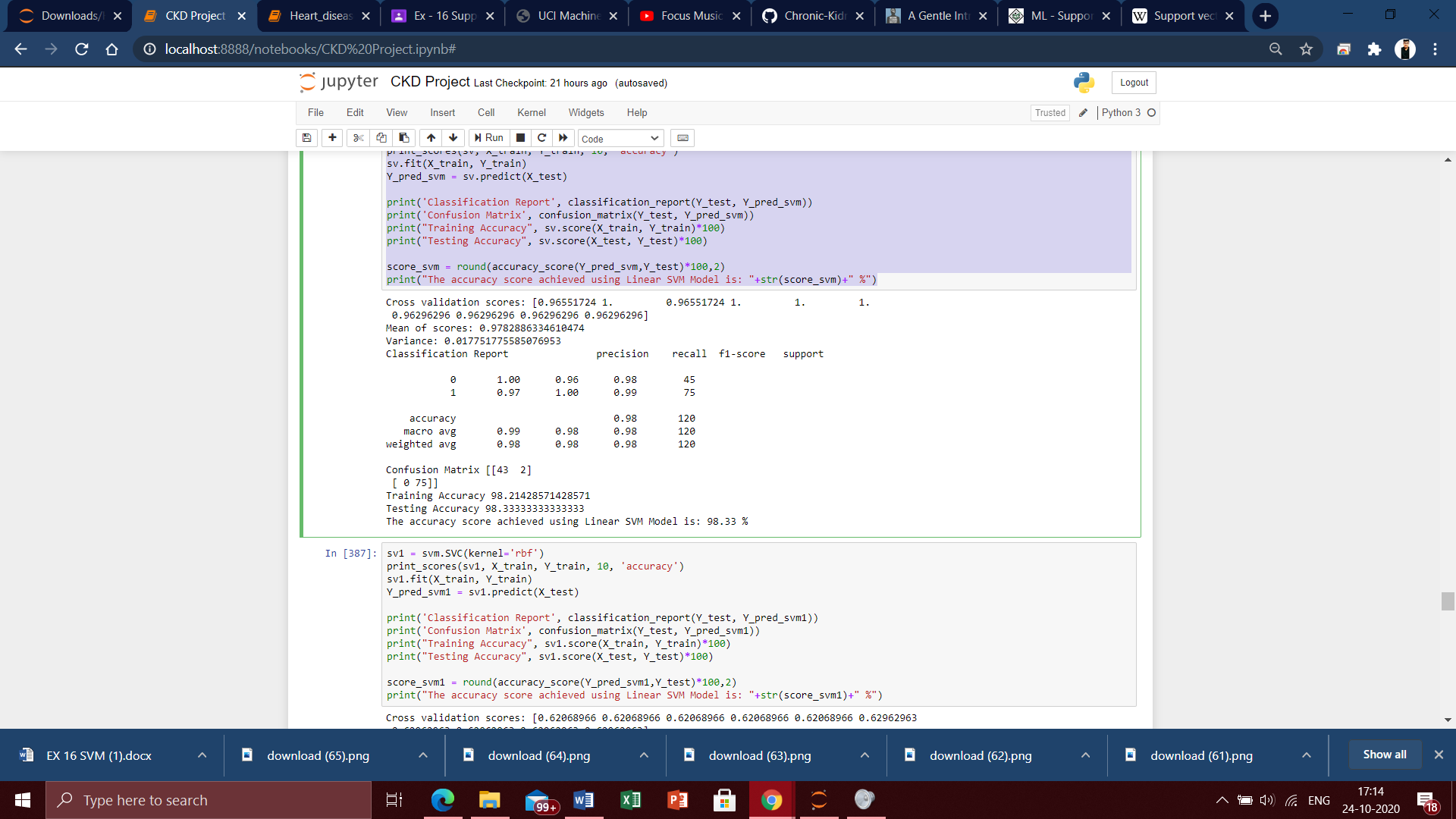
From the Naive Bayes model we have first obtained the K fold Cross validation scores for about 10 folds. Majority of the scores are 1 and there are chances that our model’s accuracy will be 100%. Next the mean of the scores we obtained that is about 0.99629, variance of about 0.01111. Here the variance is so low so there are no errors occurred in our model. Now comes the classification report which displays the precision, recall, F1 and support scores for the model.

* Precision Score for class 0 (negative) is 1.00 and for class 1 (positive) is 1.00, indicating the preciseness of the model i.e. so accurate.
* Recall value for class 0 is 1.00 and class 1 is 1.00, which describes the amount up – to which the model can predict the output.
* As the precision and recall values are similar, there are no dominance in classes.

From the confusion matrix we can see that there are 45 true positives, 0 false positive and false negative and finally 75 true negative values. The model didn’t even make a slight error and predicted the values perfectly. The training and test accuracy of the model is 100%. Finally we have come to the accuracy of the model. Surprisingly, we have the model accuracy of 100%. The model has worked well and classified the patients with and without CKD with an accuracy of 100%. The model is also not over-fitted as model optimisation techniques like K fold cross validation is done for 10 folds. So the Naive Bayes Classifier can effectively classify the patients with and without CKD precisely.

**Support Vector Machines:**

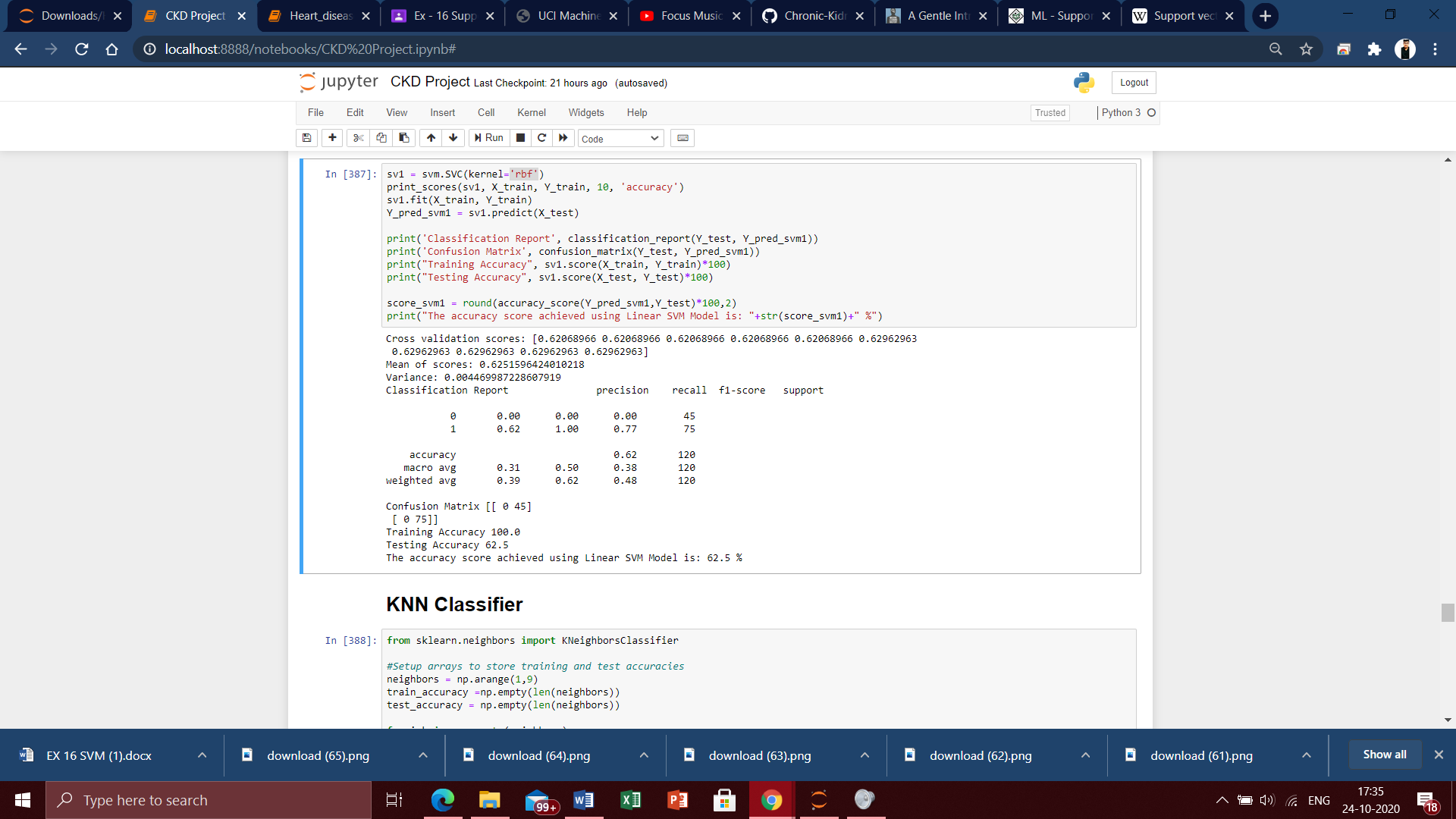
Support Vector Machine (SVM) are supervised learning models with related learning algorithms that examine data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. This is the model which took the highest time complexity to run. SVM works by mapping data to a high-dimensional feature space so that data points can be classified, even when the data are not otherwise linearly separable. Here we will be using Support vector classifiers with both Linear and Radial Basis kernel to see how this model identifies the patients with and without CKD accurately.



The linear kernel SVC fitted to predict whether the patients have CKD or not has come with good results. From the SVM model we have first obtained the K fold Cross validation scores for about 10 folds. Majority of the scores are falling in the range of 0.96 to 1, and there are chances that our model’s accuracy will be between them. Next the mean of the scores we obtained that is about 0.97828, variance of about 0.01775. Here the variance is so low so there are no errors occurred in our model. Now comes the classification report which displays the precision, recall, F1 and support scores for the model.

* Precision Score for class 0 (negative) is 1.00 and for class 1 (positive) is 0.97, indicating the preciseness of the model.
* Recall value for class 0 is 0.96 and class 1 is 1.00, which describes the amount up – to which the model can predict the output.
* As the precision and recall values are slightly different from one another, so there are chances for dominance in classes.

From the confusion matrix we can see that there are 43 true positives, 2 false positives, 0 false negative and finally 75 true negative values. The model made an error here giving the prediction for false positives. And that is why the accuracy has reduced comparing to the other models. The training and test accuracy of the model is 98.21% and 98.33% respectively. Finally we have come to the accuracy of the model. Here we have the model accuracy of 98.33%. The model has worked well and classified the patients with and without CKD with a slight error with an accuracy of 98.33%. Errors in the medical fields is not acceptable, even though the model tried its best after optimization too. The model is also not over-fitted as model optimisation techniques like K fold cross validation is done for 10 folds. Thus, the linear kernel SVC has made a good classification with only a slightest error.



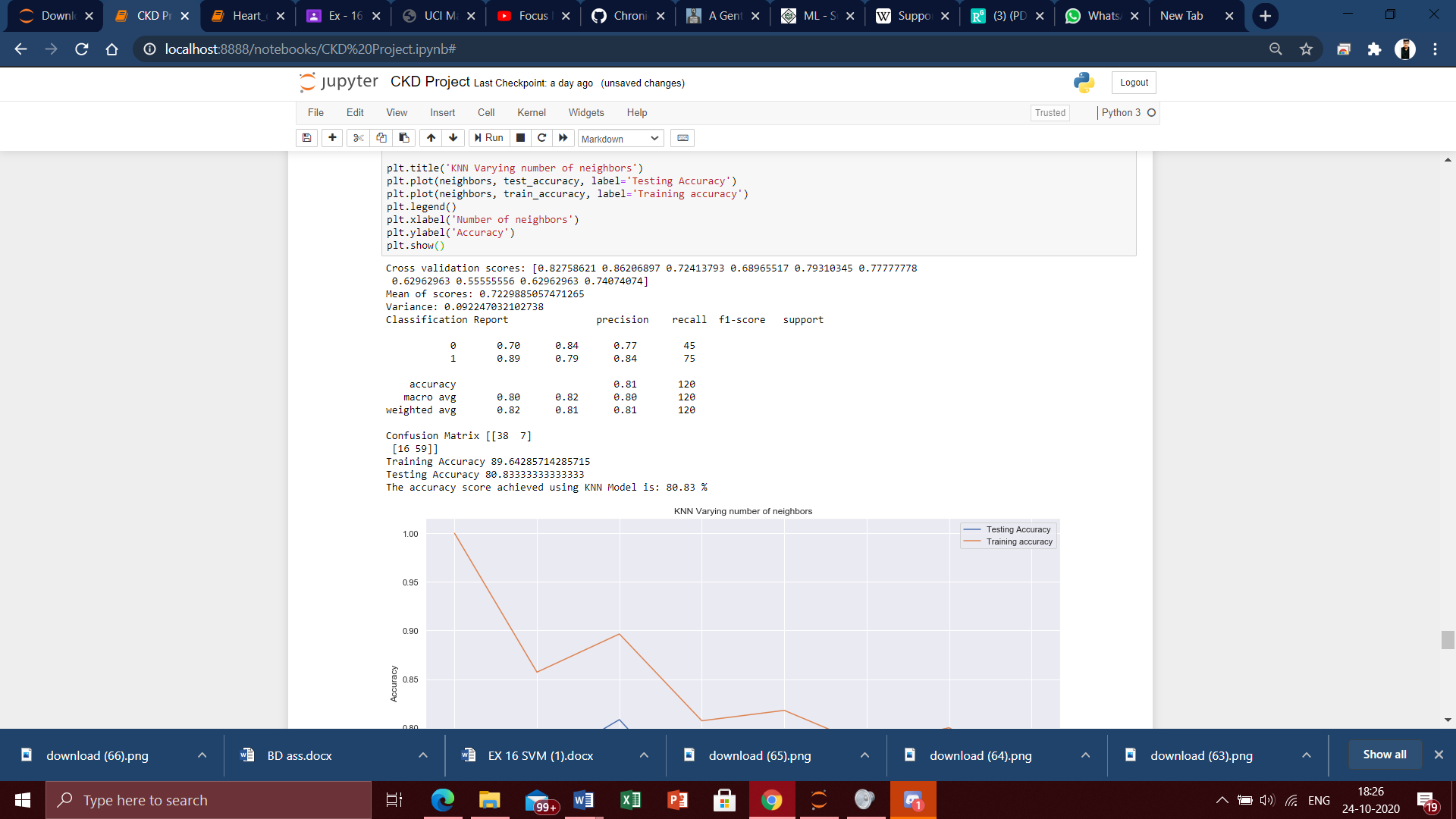
The Radial Basis kernel SVC fitted to predict whether the patients have CKD or not has come with worst results. From the SVM model we have first obtained the K fold Cross validation scores for about 10 folds. Majority of the scores are fall on 0.62, and there are chances that our model’s accuracy will drop down to sixties. Next the mean of the scores we obtained is about 0.62515 variance of about 0.00446. Now comes the classification report which displays the precision, recall, F1 and support scores for the model.

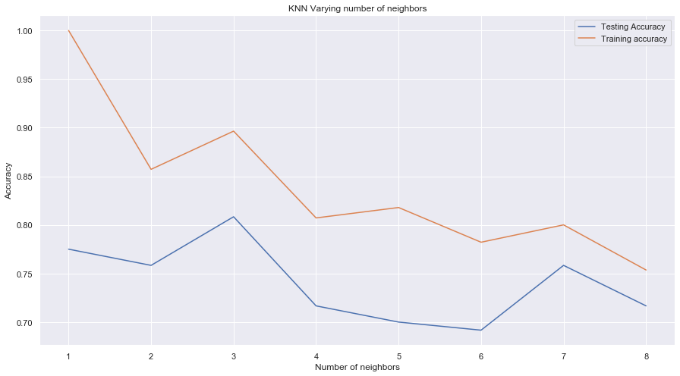
* Precision Score for class 0 (negative) is 0.00 and for class 1 (positive) is 0.62, indicating the preciseness of the model which is very worse.
* Recall value for class 0 is 0 and class 1 is 1.00, which describes the amount up – to which the model can predict the output. The positive class’s recall is fine but the negative’s recall is 0 which is a poor recall.
* As the precision and recall values are totally different high dominance of classes will occur

From the confusion matrix we can see that there are 0 true positive, 45 false positives, and 0 false negative and finally 75 true negative values. The model made a huge error here giving the prediction for true and false positives, which are the deciding factors of the classification. And that is why the accuracy has reduced comparing to the other models. The training and test accuracy of the model is 100% and 62.5% respectively. Finally we have come to the accuracy of the model. Here we have the model accuracy of 62.5%. The model has worked very poor and classified the patients with and without CKD with more errors with an accuracy of 62.5%. Errors in the medical fields is not acceptable, even though the model tried its best after optimization too. The model is also not under -fit as model optimisation techniques like K fold cross validation is done for 10 folds. Thus, the Radial Bases kernel SVC has made the worst classification of patients with and without the renal disease.

**KNN Classifier:**

K-Nearest Neighbours (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures). Classification is done by a majority vote to its neighbours. Classification is done by a majority vote of neighbours. If K = 1, then the class is single nearest neighbour. In a common weighting scheme, individual neighbour is assigned to a weight of 1/d if d is the distance to the neighbour. The shortest distance between any two neighbours is always a straight line and the distance is known as Euclidean distance. In the same way the KNN classifies whether the patient has CKD or not by considering the attributes involved in classification.





**Inference:**

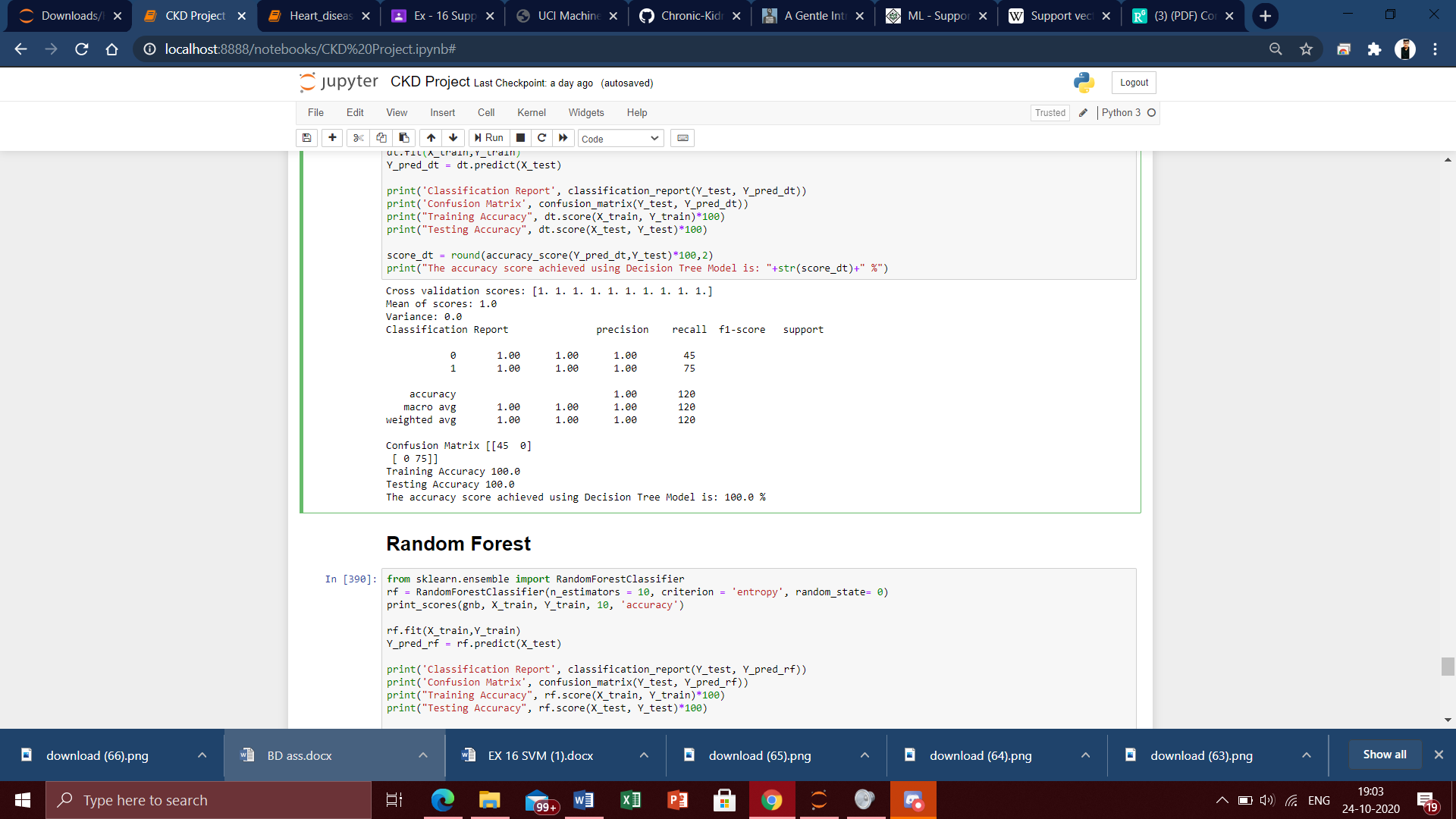
The KNN Classifier fitted to predict whether the patients have CKD or not has come with fair results. From the KNN model we have first obtained the K fold Cross validation scores for about 10 folds. Majority of the scores are fall on range between 0.55-0.82, and there are chances that our model’s accuracy will drop down to eighties. Next the mean of the scores we obtained is about 0.72998, variance of about 0.0944. Here the variance is so low so there are only few errors occurred in our model. From the graph we can infer that in the testing accuracy there is a peak when the k value is 3. So we consider the k value is 3 for proceeding the model. Now comes the classification report which displays the precision, recall, F1 and support scores for the model.

* Precision Score for class 0 (negative) is 0.70 and for class 1 (positive) is 0.89, indicating the preciseness of the model which is fair.
* Recall value for class 0 is 0.84 and class 1 is 0.79, which describes the amount up – to which the model can predict the output. The positive class and negative class’s recalls are quite fine.
* As the precision and recall values are quite different, the dominance of classes may occur at some places.

From the confusion matrix we can see that there are 38 true positives, 7 false positives, 16 false negatives and finally 59 true negative values. The model made small errors on all the instances of the confusion matrix. And that is why the accuracy has reduced to 80s when compared to the other models. The training and test accuracy of the model is 89.64% and 80.83% respectively. Finally we have come to the accuracy of the model. Here we have the model accuracy of 80.83%. The model has worked fair and classified the patients with and without CKD with slight errors with an accuracy of 80.83%. Errors in the medical fields is not acceptable, even though the model tried its best after optimization too. The model is also not under -fitted as model optimisation techniques like K fold cross validation is done for 10 folds. Thus, KNN Classifier has made a fair classification of patients with and without the CKD.

**Decision Tree Classifier:**

Decision Tree is one of the classification algorithms, which is used to solve regression and classification problems. The general objective of using Decision Tree is to create a model that predicts classes or values of target variables by generating decision rules derived from training data sets. Decision tree algorithm follows a tree structure with roots, branches and leaves. The attributes of decision making are the internal nodes and class labels are represented as leaf nodes. Decision Tree algorithm is easy to understand compared with other classification algorithms. Let us see how the decision tree classifies the patients with and without CKD.



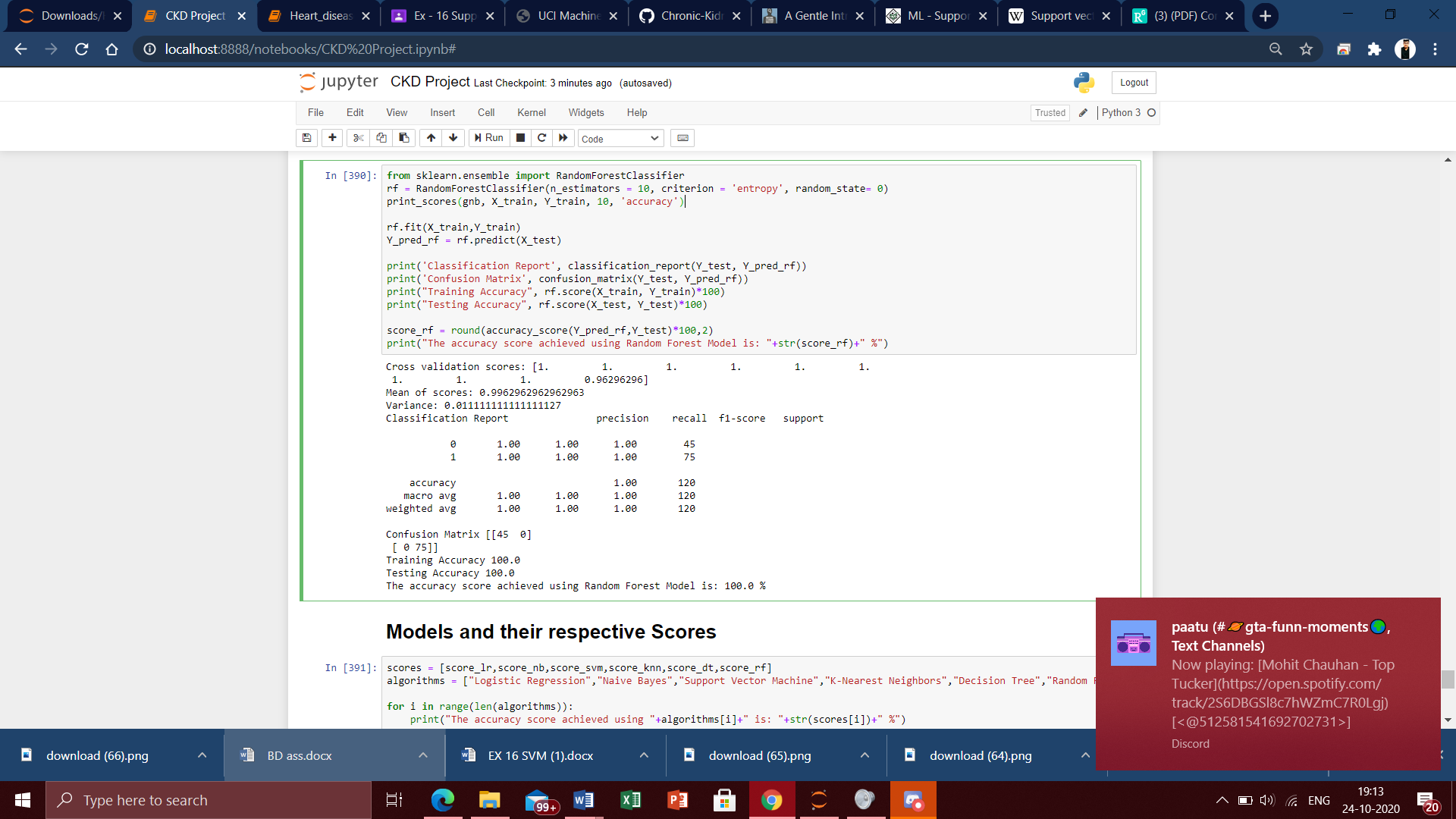
Entropy criterion is used to build the decision tree classifier. It uses information gain to decide which attribute it should choose to branch the tree. From the decision tree model we have first obtained the K fold Cross validation scores for about 10 folds. All the scores are 1 and there are chances that our model’s accuracy will be 100%. Next the mean of the scores we obtained that is about 1, variance of about 0. Here the variance is 0 so there are no errors occurred in our model and is perfectly fitted. Now comes the classification report which displays the precision, recall, F1 and support scores for the model.

* Precision Score for class 0 (negative) is 1.00 and for class 1 (positive) is 1.00, indicating the preciseness of the model which is perfect.
* Recall value for class 0 is 1.00 and class 1 is 1.00, which describes the amount up – to which the model can predict the output.
* As the precision and recall values are similar, there are no dominance in classes.

From the confusion matrix we can clearly see that there are 45 true positives, 0 false positive and false negative and finally 75 true negative values. The model didn’t even make a slight error and predicted the values perfectly. The training and test accuracy of the model are 100% and 100% respectively. Finally we have come to the accuracy of the model. Surprisingly, we have the model accuracy of 100%. The model has worked well and classified the patients with and without CKD with an accuracy of 100%. The model is also not over-fitted as model optimisation techniques like K fold cross validation is done for 10 folds. So the decision tree model can effectively classify the patients without even a slightest error.

**Random Forest Classifier:**

Random forest algorithm constructs multiple decision trees to act as an ensemble of classification and regression process. A number of decision trees are constructed using a random subsets of the training data sets. A large collection of decision trees provide higher accuracy of results. The runtime of the algorithm is comparatively fast and also accommodates missing data. Random forest randomizes the algorithm and not the training data set. The decision class is the mode of classes generated by decision trees. Here we will be using Random forest classifier with to see how this model identifies the patients with and without CKD accurately.



Entropy criterion is used to build the random forest classifier. It uses information gain to decide which attribute it should choose to branch the tree. From the Random forest model we have first obtained the K fold Cross validation scores for about 10 folds. All the scores are 1 except the last fold, and there are chances that our model’s accuracy will be 100%. Next the mean of the scores we obtained that is about 0.99, variance of about 0.01. Here the variance is very low, so there are no errors occurred in our model and is perfectly fitted. Now comes the classification report which displays the precision, recall, F1 and support scores for the model.

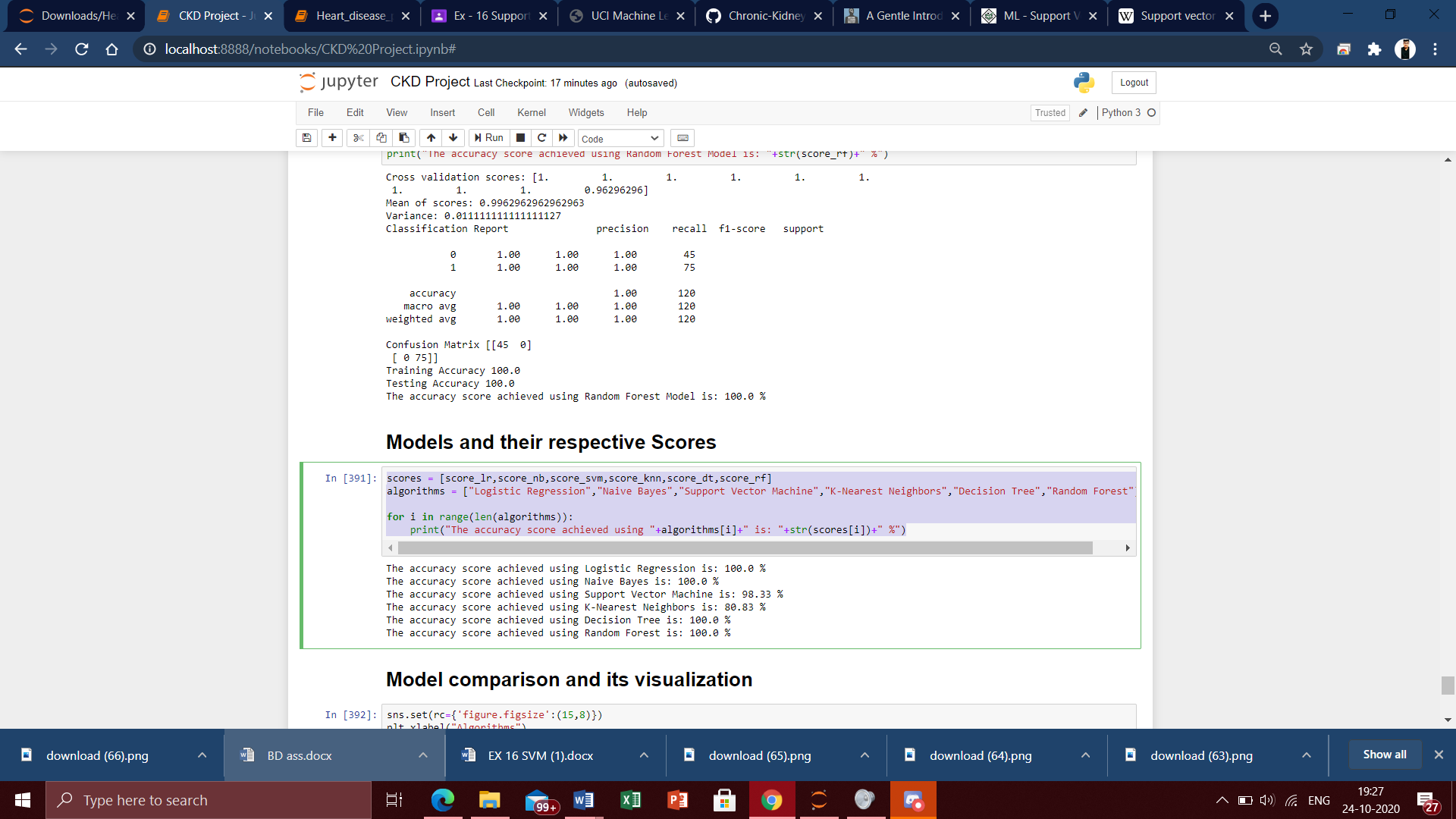
* Precision Score for class 0 (negative) is 1.00 and for class 1 (positive) is 1.00, indicating the preciseness of the model.
* Recall value for class 0 is 1.00 and class 1 is 1.00, which describes the amount up – to which the model can predict the output.
* As the precision and recall values are similar, there are no dominance in classes.

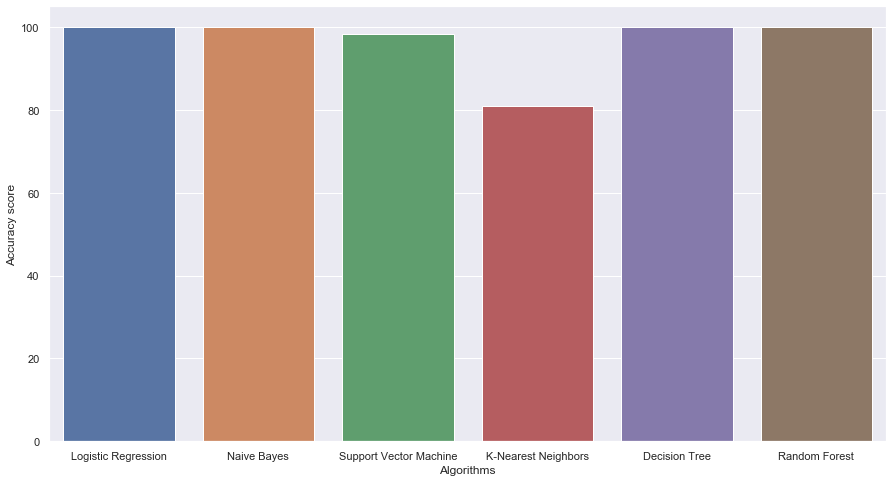
From the confusion matrix we can clearly see that there are 45 true positives, 0 false positive and false negative and finally 75 true negative values. The model didn’t even make a slight error and predicted the values perfectly. The training and test accuracy of the model are 100% and 100% respectively. Finally we have come to the accuracy of the model. Surprisingly, we have the model accuracy of 100%. The model has worked well and classified the patients with and without CKD with an accuracy of 100%. The model is also not over-fitted as model optimisation techniques like K fold cross validation is done for 10 folds. So the Random forest model can effectively classify the patients without even a slightest error.

**Model comparisons and their respective scores with visualization:**

Now all the classifiers have been fitted completely to the chronic kidney disease dataset. Let us clearly see accuracy of all the models, which can help us to infer the best model.

**Output:**





**Inference from the models constructed:**

From the models constructed above we can infer that the models with the highest accuracy of 100% is Logistic Regression, Naive Bayes Classifier, Decision Tree and Random Forest. The model with an accuracy lesser than 100% and also a good model is Support Vector Classifier with an accuracy of 98.33%. The model which had a fair performance is KNN Classifier with an accuracy of 80.83%. The model with the worst accuracy is the SVM model with Radial Basis kernel with an accuracy of 62.5%. Deciding the best model out the 4 models is quite a competitive task. But by looking into the overall accuracy, precision, recall, mean of scores and variance the **DECISION TREE MODEL** is the winner here. The reason behind the statement is the variance and mean of scores is also 0 and 1 respectively, which is perfect!

**Conclusion:**

The results of these predictions can be added to the domain of healthcare and can be used for providing suggestions in the domain by making it easy for healthcare professionals in diagnosis and treatment of patients as well as for identifying relationships within diseases that patients have. Future work should mainly focus on implementing more big data oriented tools and techniques which makes the process much faster and effective. The greatest challenge in healthcare domain is the data, provided enough and appropriate data is available, many applications can be implemented which take the healthcare industry to an advanced level.

**Reference:**

[1]. Gunarathne W.H.S.D,Perera K.D.M, Kahandawaarachchi K.A.D.C.P, “Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)”,2017 IEEE 17th International Conference on Bioinformatics and Bioengineering.

[2]. S.Ramya, Dr. N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer and Communication Engineering,Vol. 4, Issue 1, January 2016.

[3]. S.Dilli Arasu and Dr. R.Thirumalaiselvi, “Review of Chronic Kidney Disease based on Data Mining Techniques”,International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13498-13505

[4]. L. Rubini, “Early stage of chronic kidney disease UCI machine learning repository,”2015.

[5]. S. A. Shinde and P. R. Rajeswari, “Intelligent health risk prediction systems using machine learning : a review,” IJET, vol. 7, no. 3, pp. 1019– 1023, 2018.

[6]. Himanshu Sharma,M A Rizvi,”Prediction of Heart Disease using Machine Learning Algorithms: A Survey”,International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169,Volume: 5 Issue: 8.