**CS 634 – Data Mining**

**Project Report**

**Topic – Supervised Data Mining (Classification)**

**Name – Tanish Bugnait**

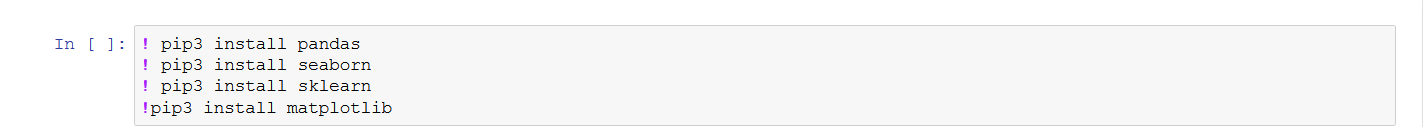
**NJIT ID – 31518500**

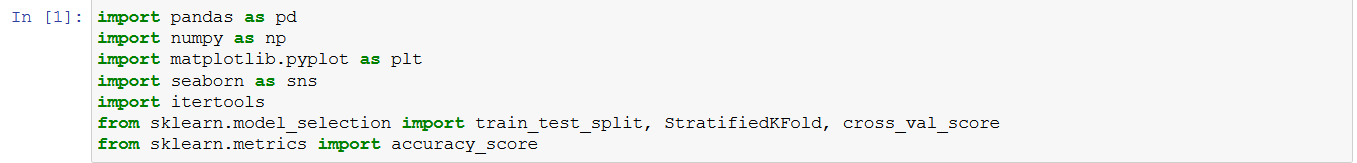
The following project demonstrates Supervised Data Mining (Classification). I had selected option 1 and implemented the following 3 evaluating classifiers:

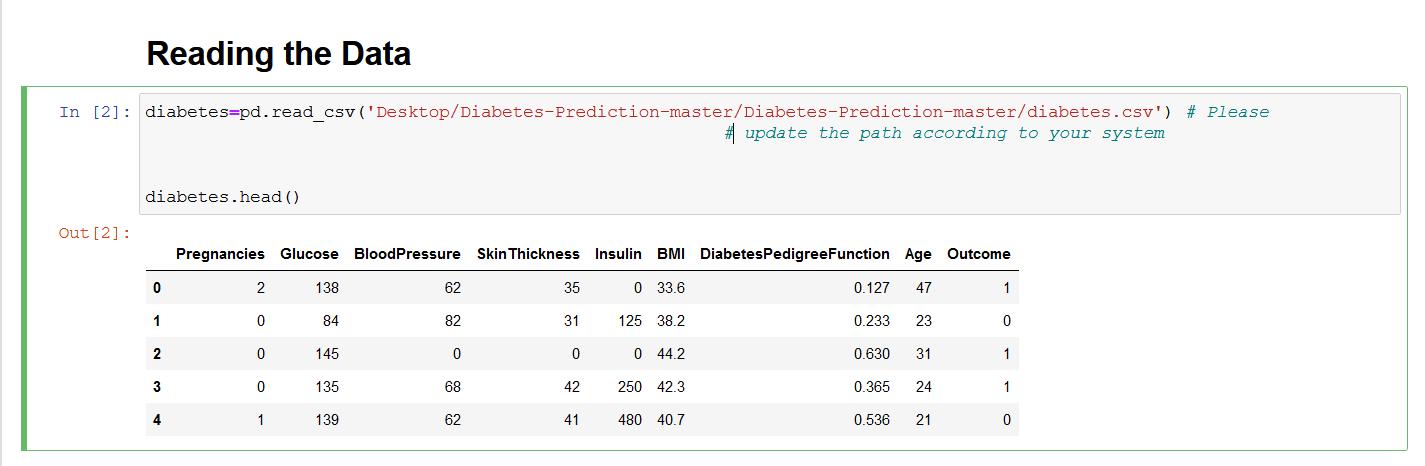
* Random Forest
* Naïve Bayes
* KNN (K- Nearest Neighbors)

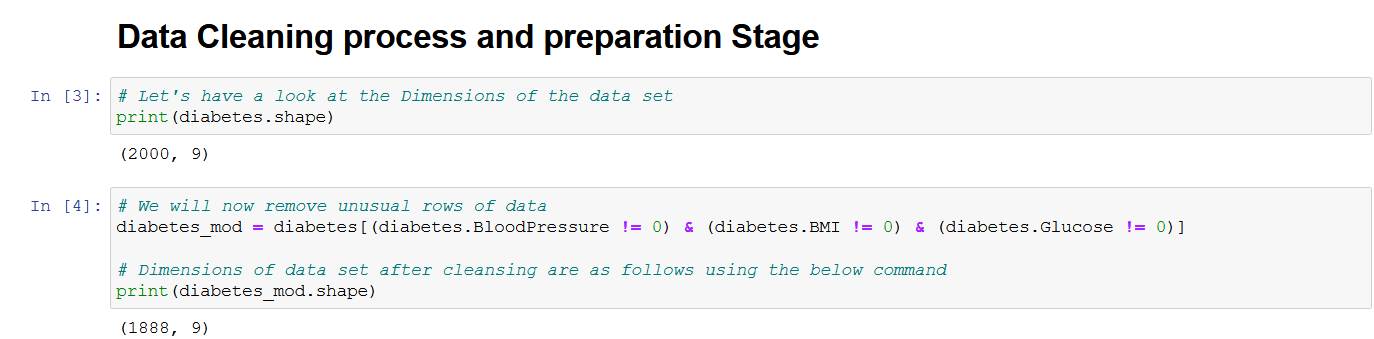
I have calculated all possible metrics that were taught in the class using the formulas given in the PPT. The Dataset that I have chosen is for Diabetes prediction and the code showcases a step-by-step implementation for all the parameters that was asked to be performed.

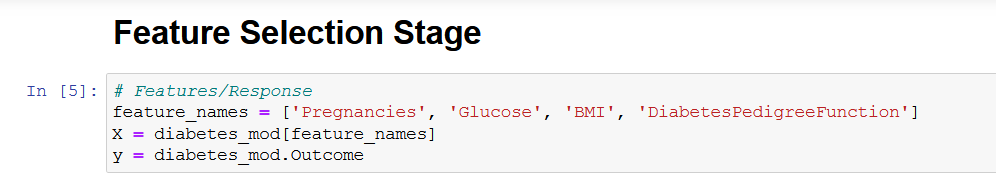
**Implementation of the code**

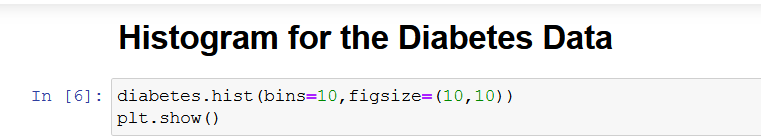


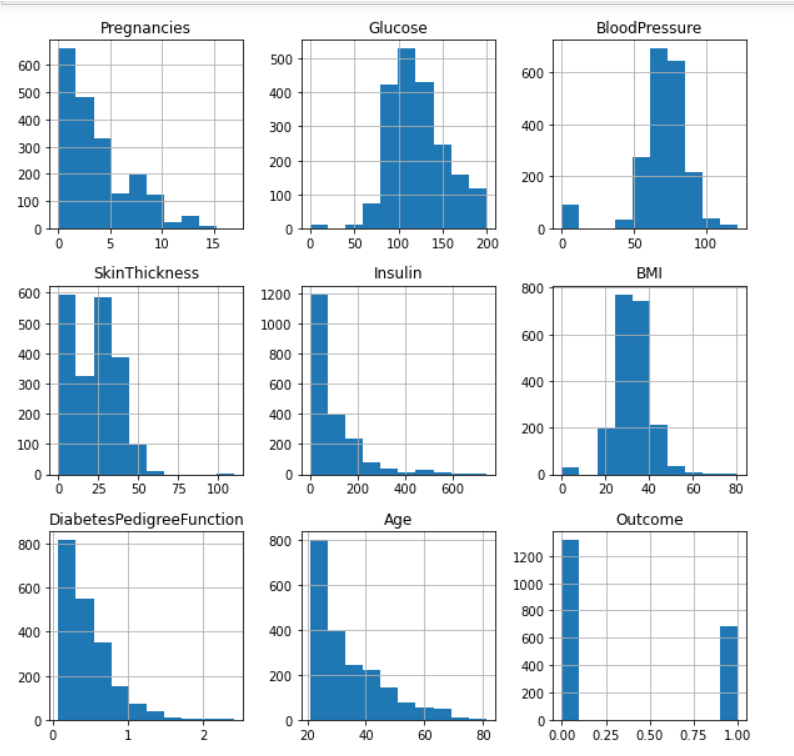


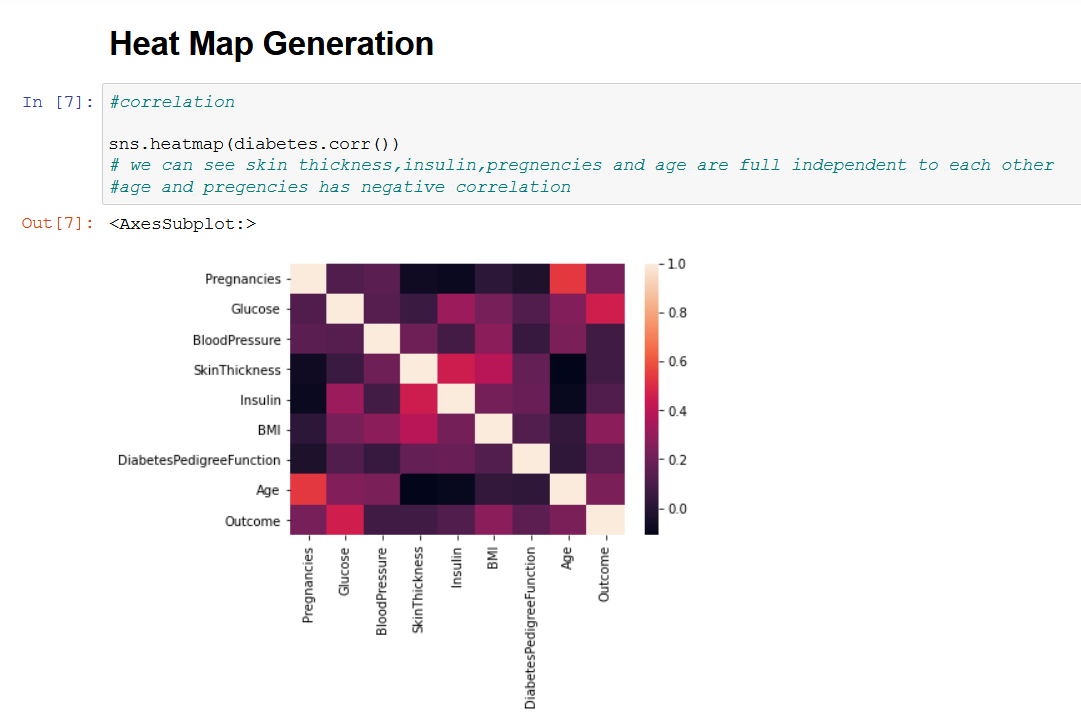


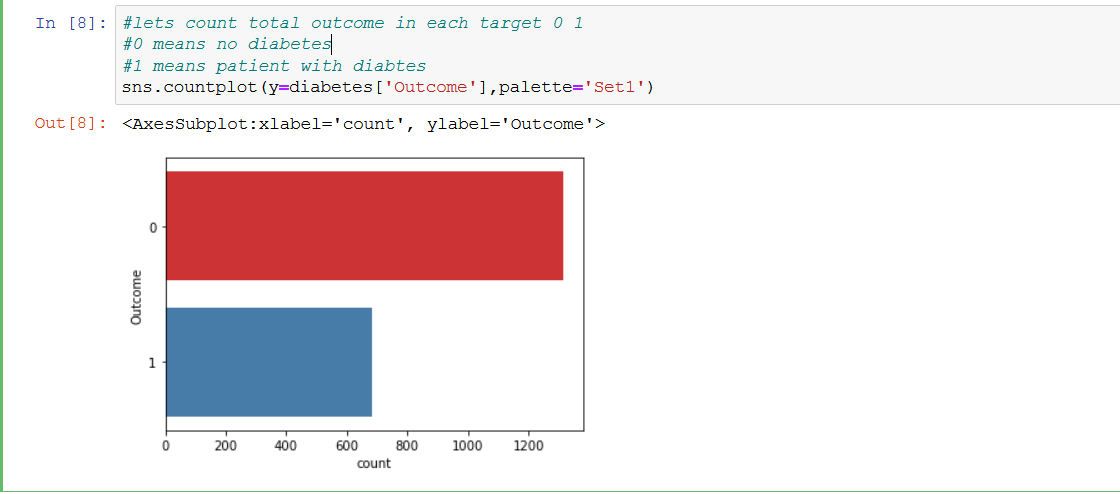


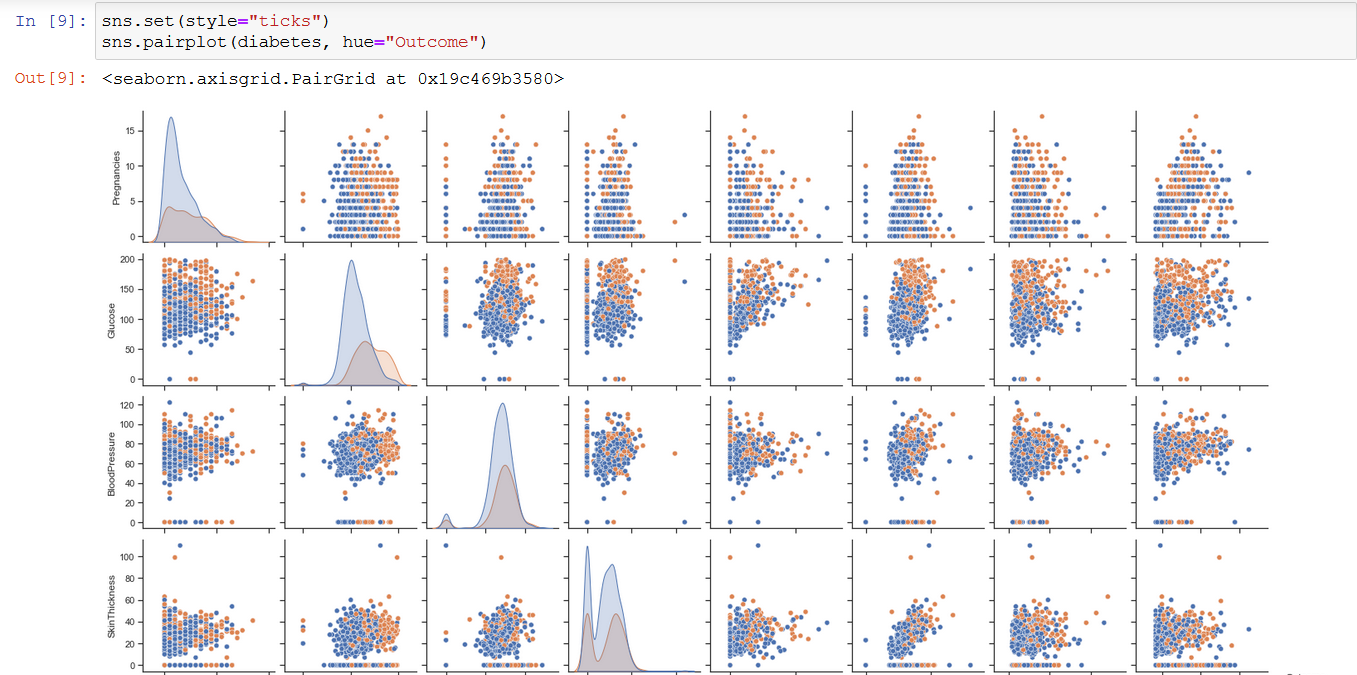


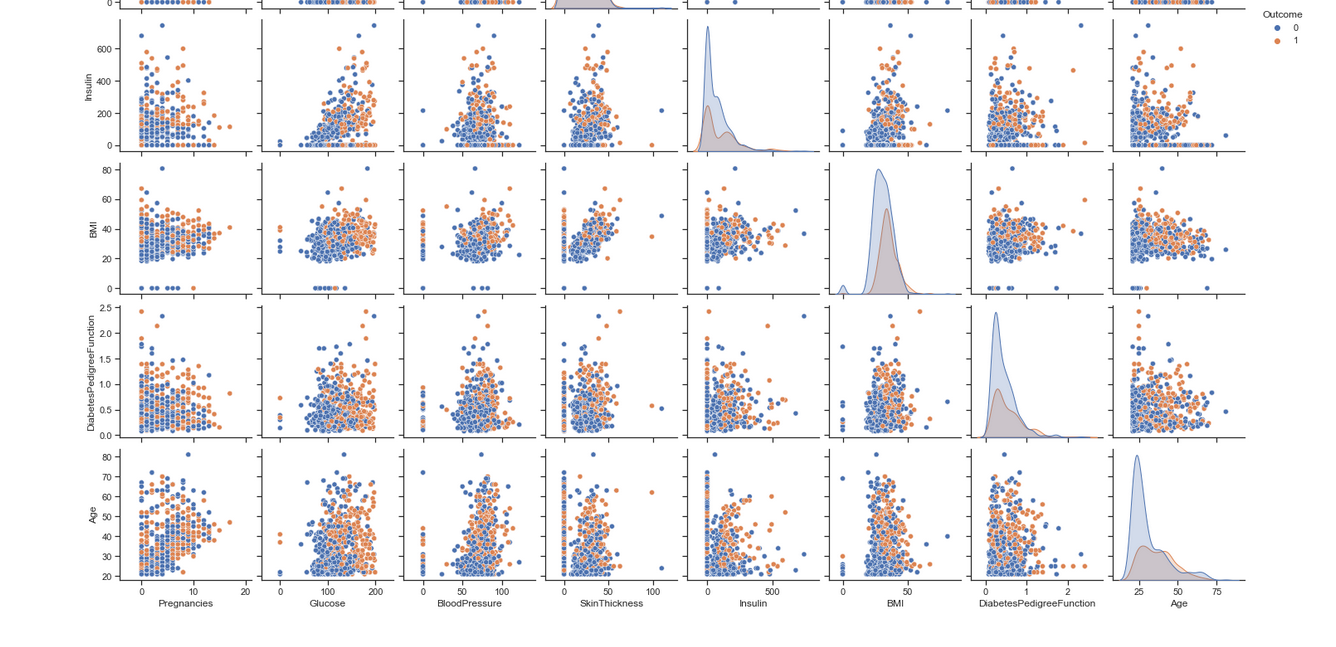


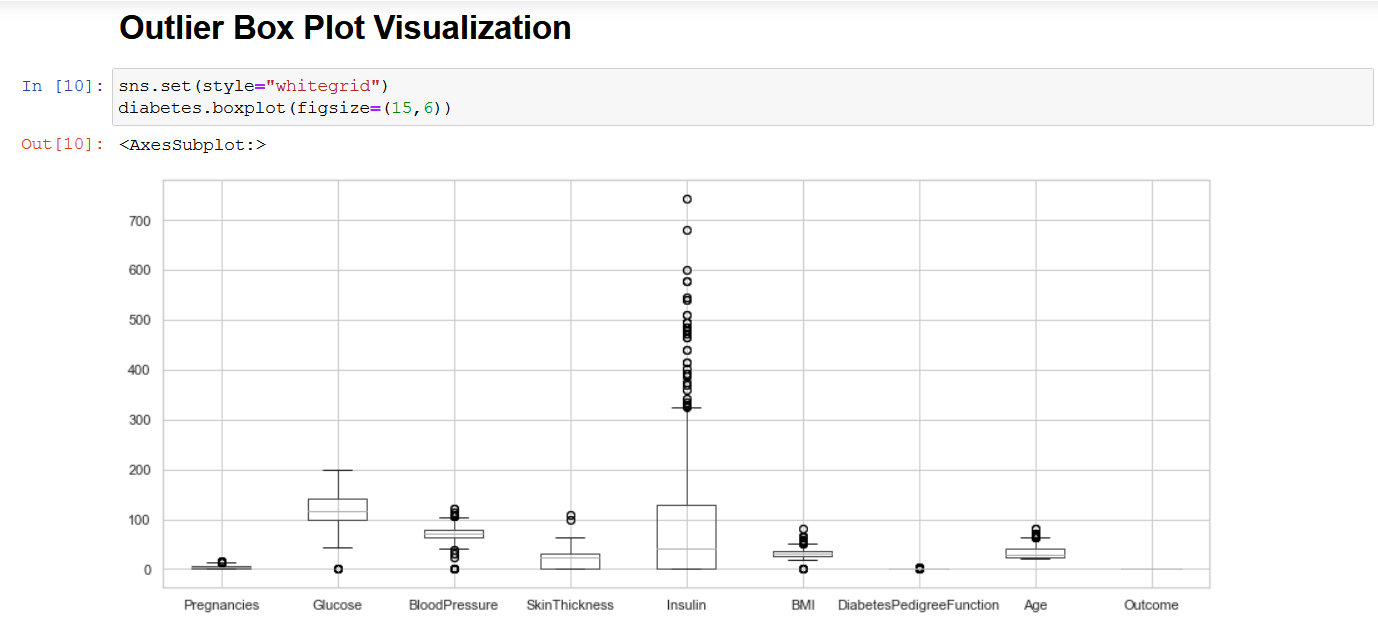






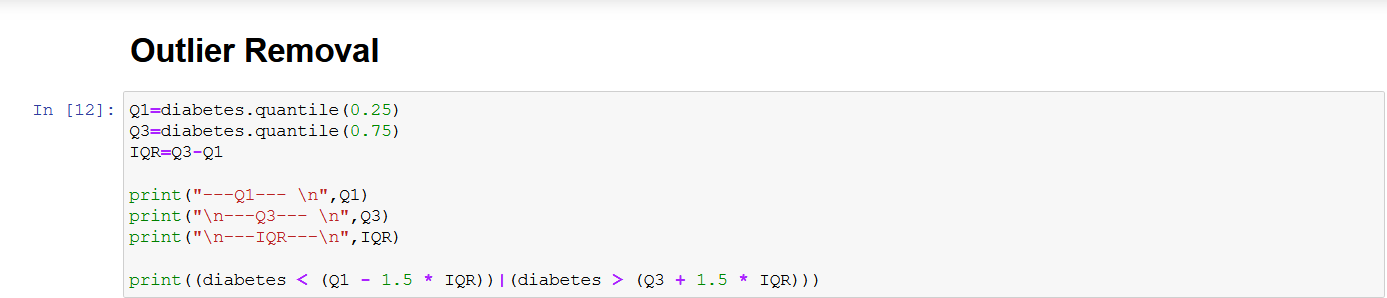


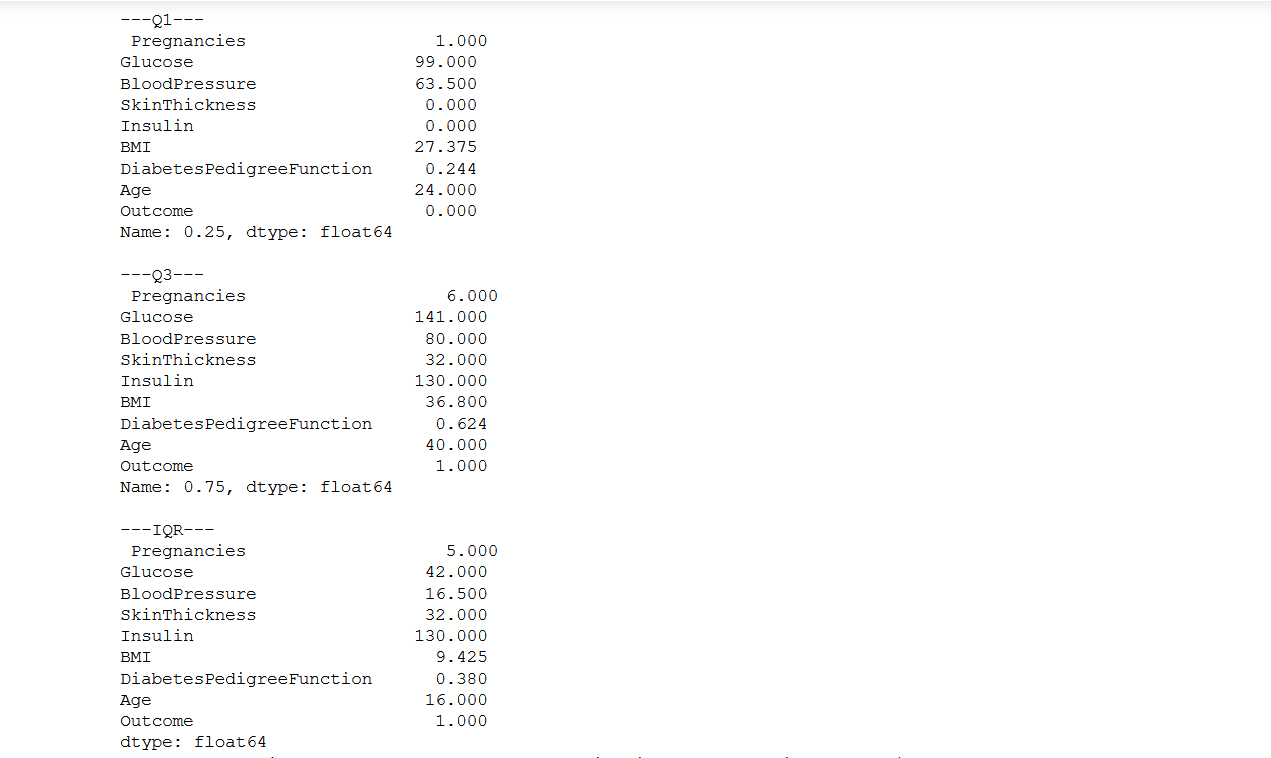


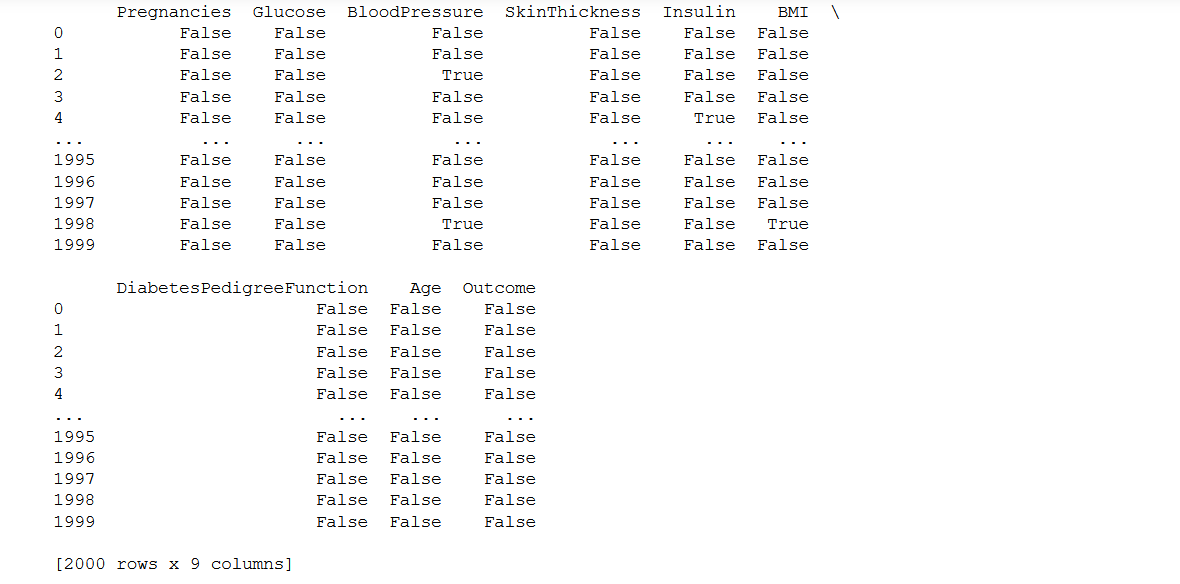


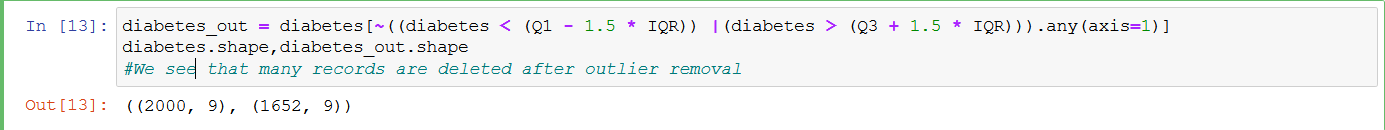




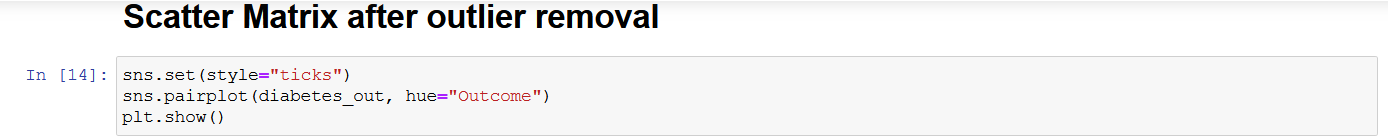


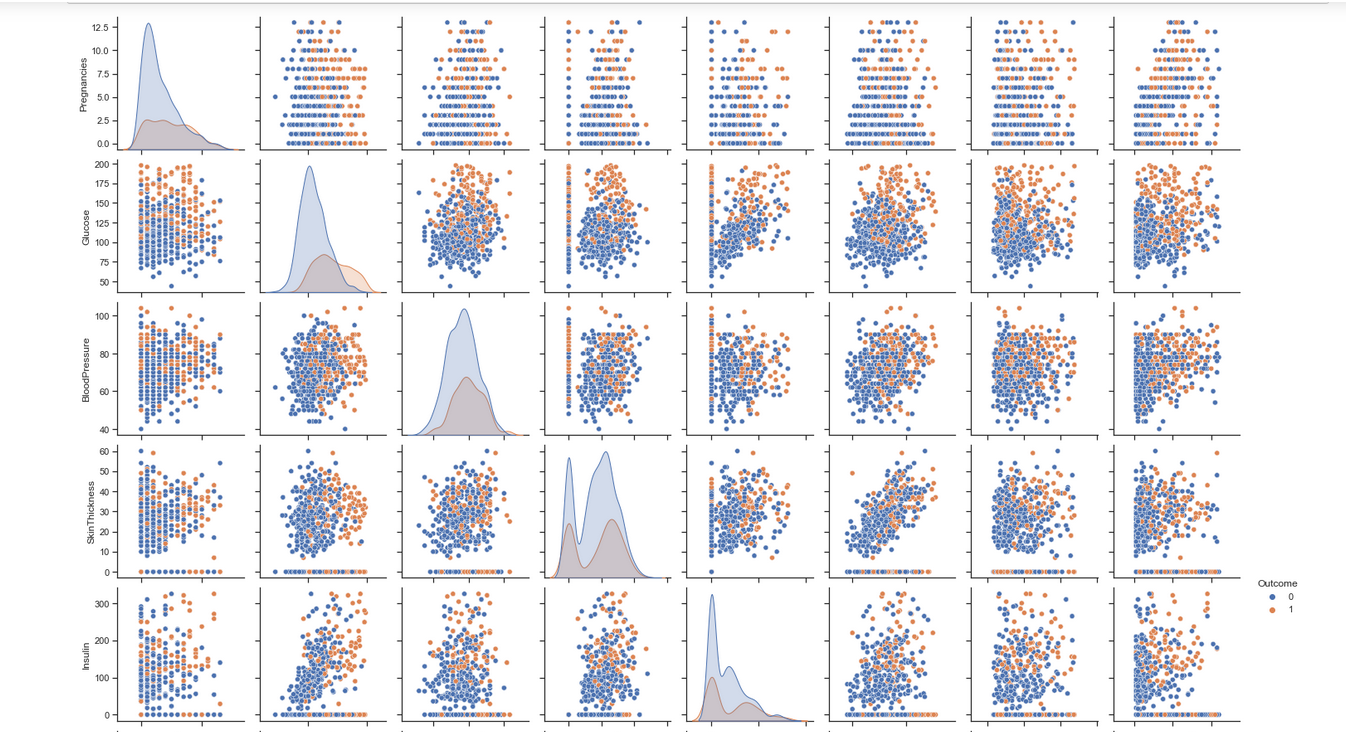


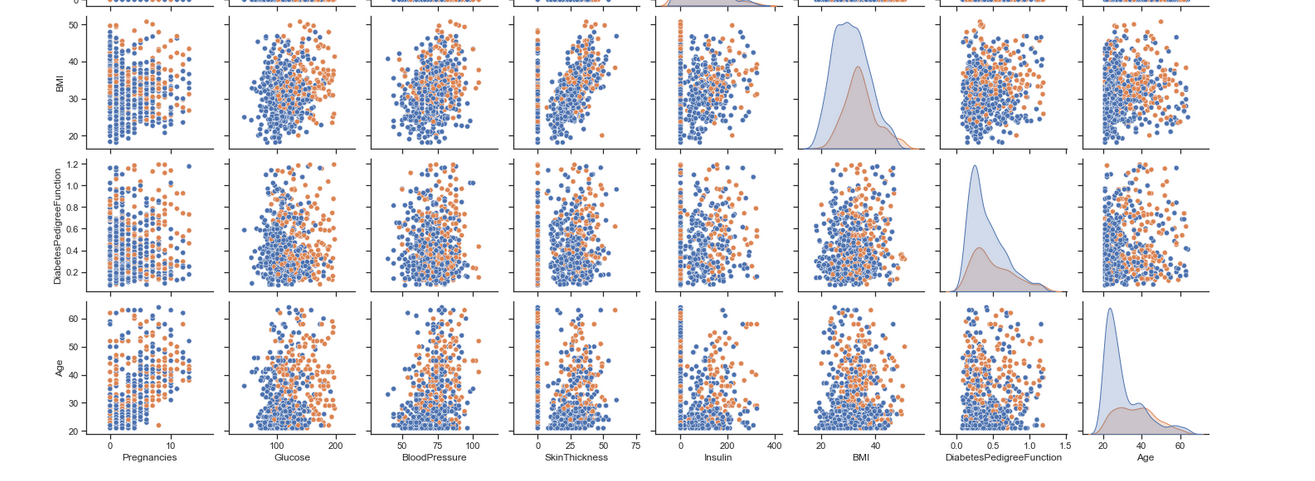




Now that the Outliers have been removed, we will have a look at the newly generated Scatter Matrix.



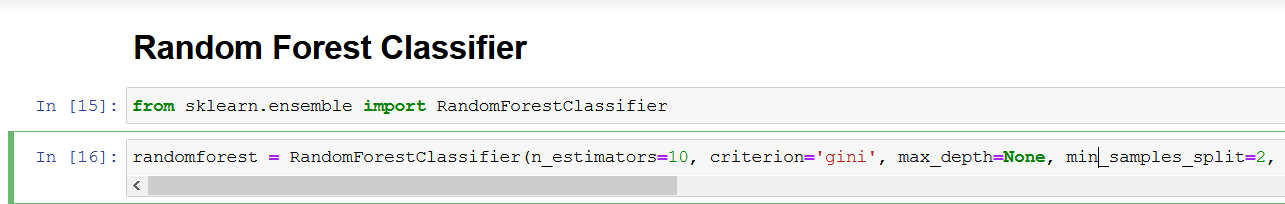


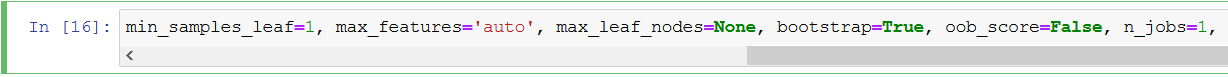


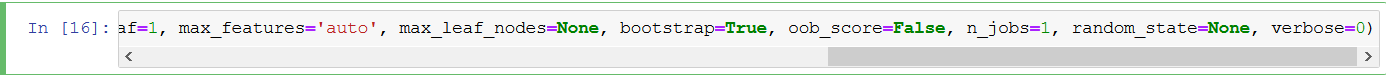
Now, it’s time to implement the 3 different classifiers and calculate different metrics as mentioned in the PPT.

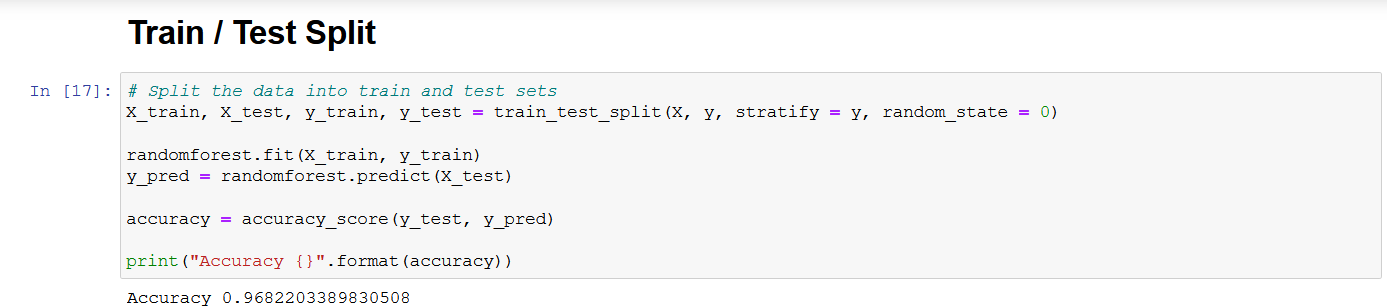
We will first have a look at the Random Forest Classifier.

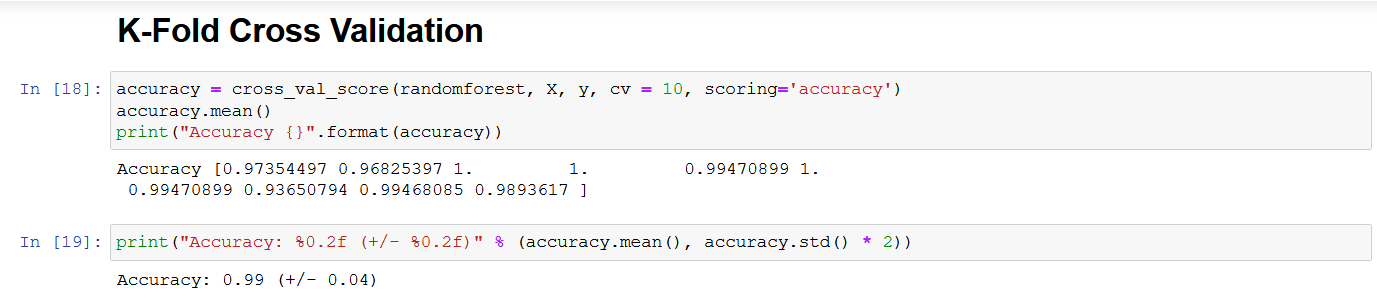
1. **Random Forest Classifier**



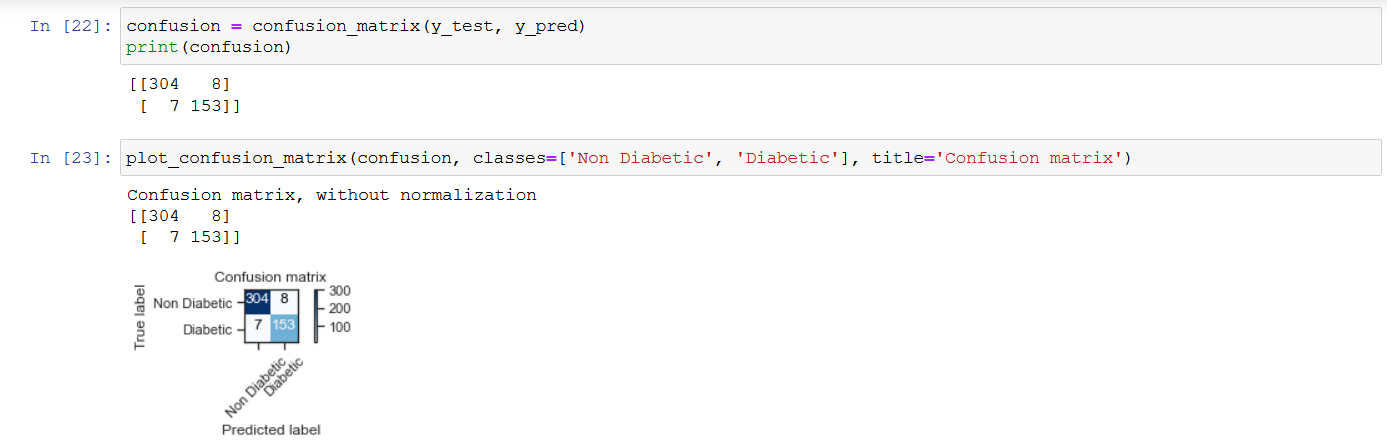


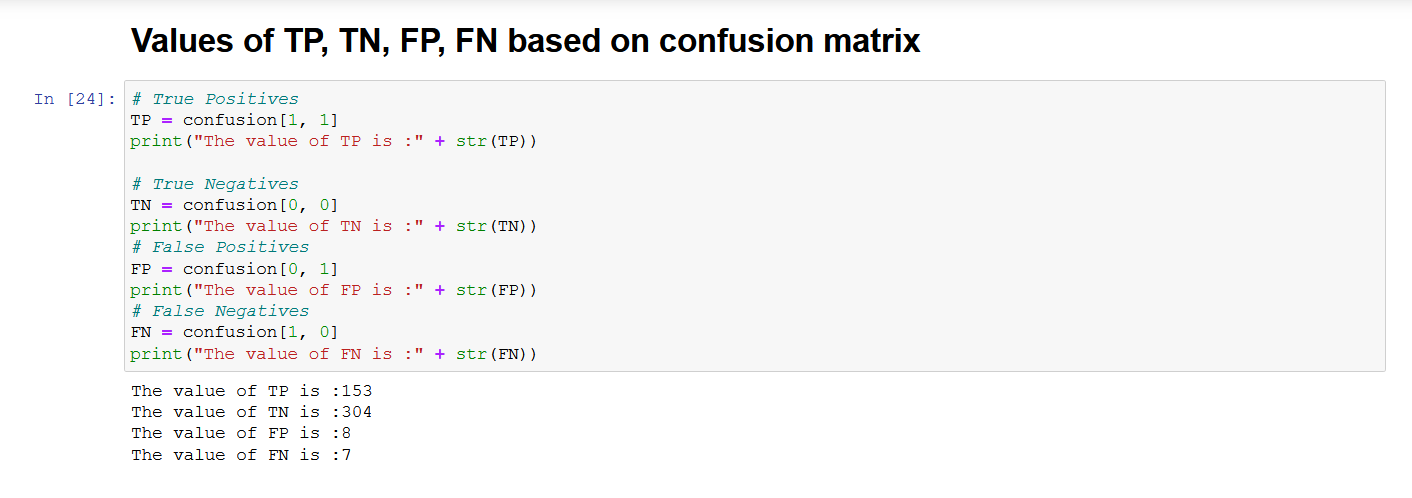


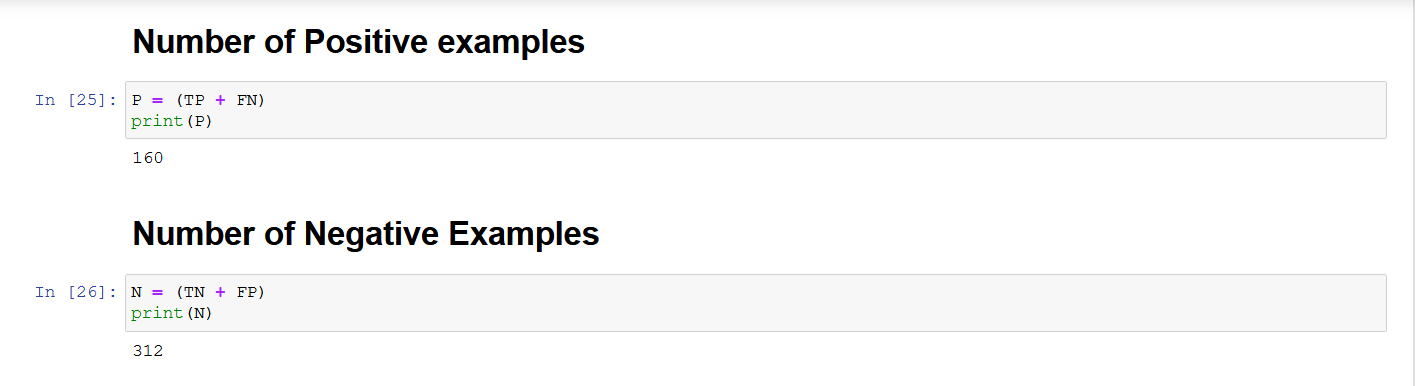






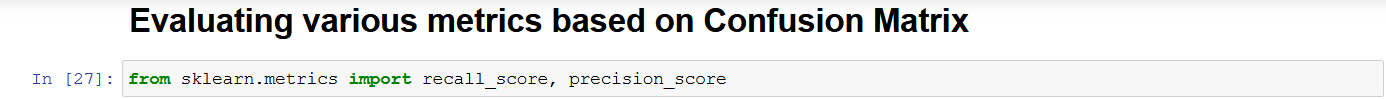




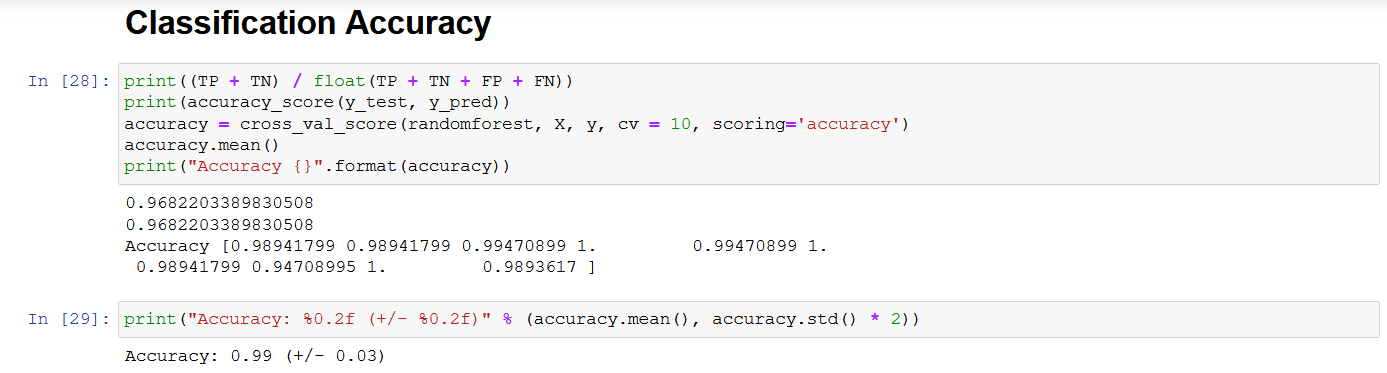


Up until now I have imported the necessary libraries required to run my Random Forest Classifier with the help of scikit learn and also made the confusion matrix.

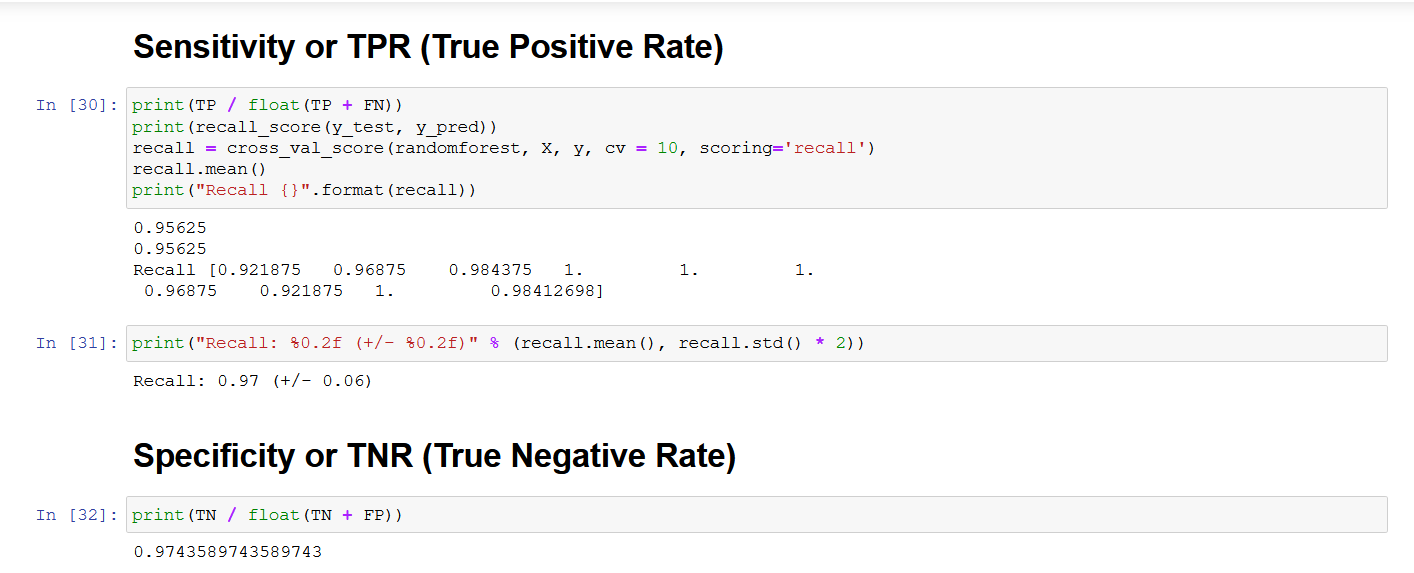
Now, it’s time to calculate the various metrics values by using the given formula.



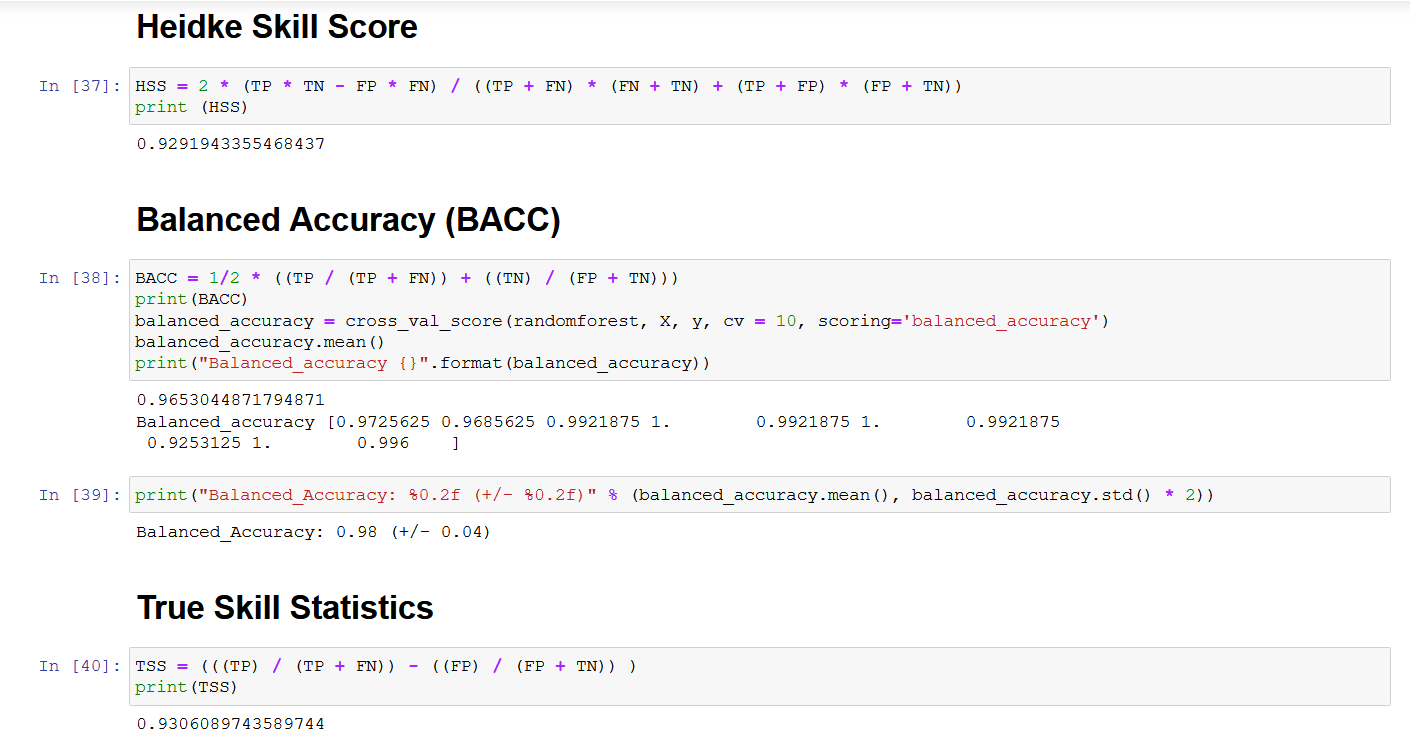
At the end of the classifier, I have provided a tabular format showcasing the different values for 10 Folds and also the average for each metric.



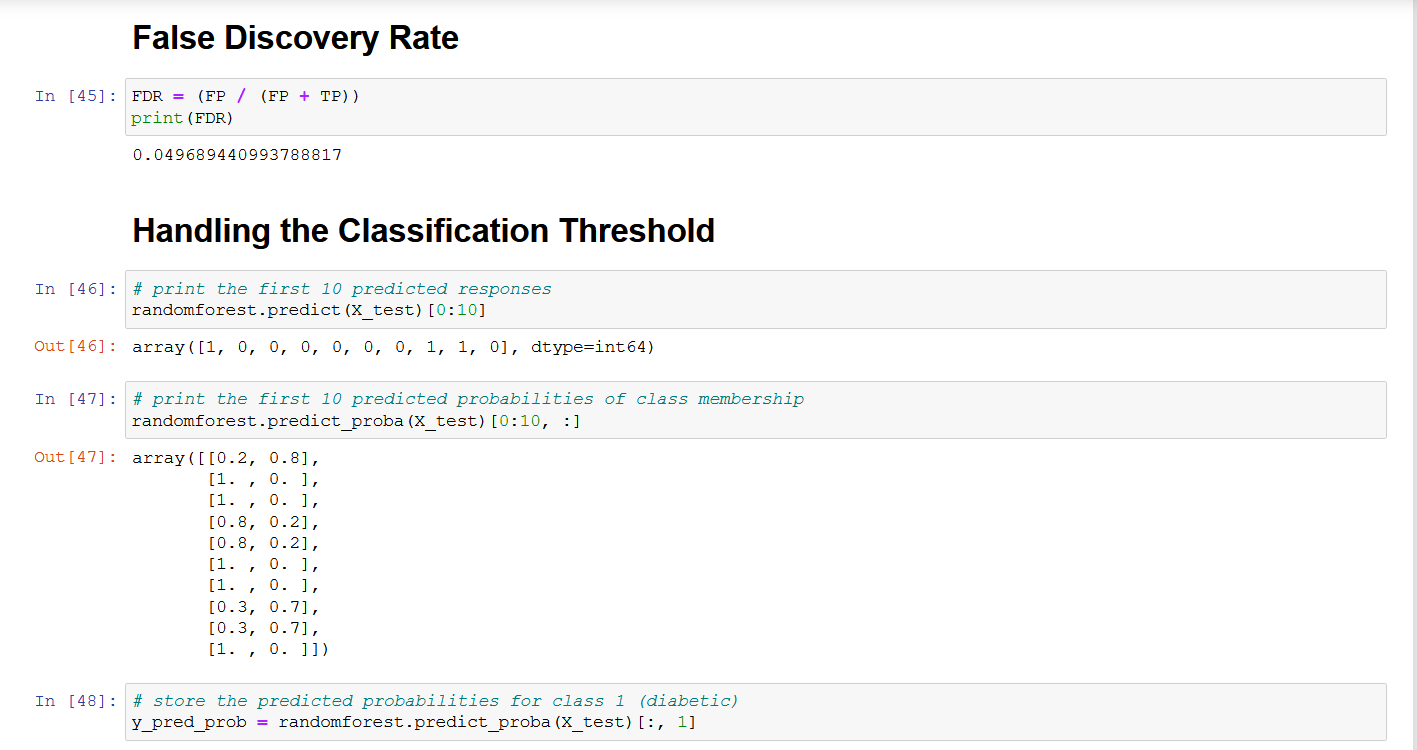
We see 10 different values of Accuracy for 10 different folds along with the mean for the same. The tabular format provided at the end of this classifier would give a clear understanding of the output.

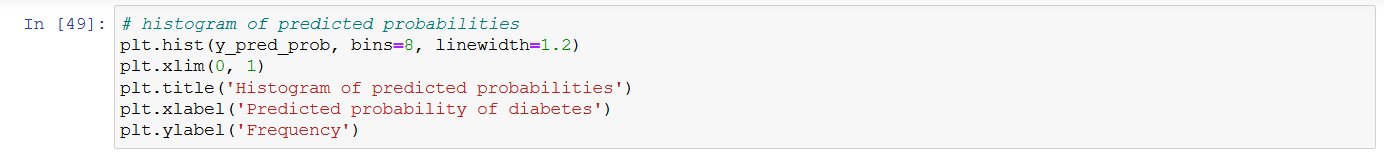


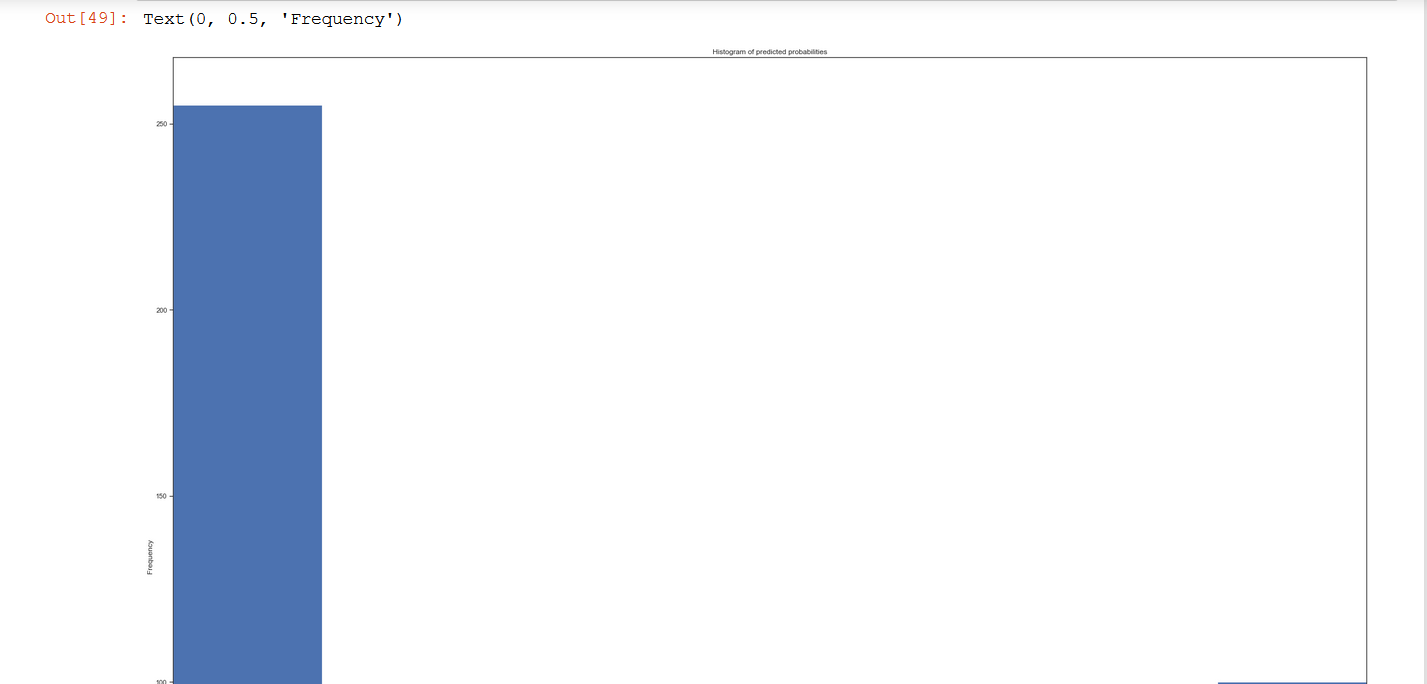


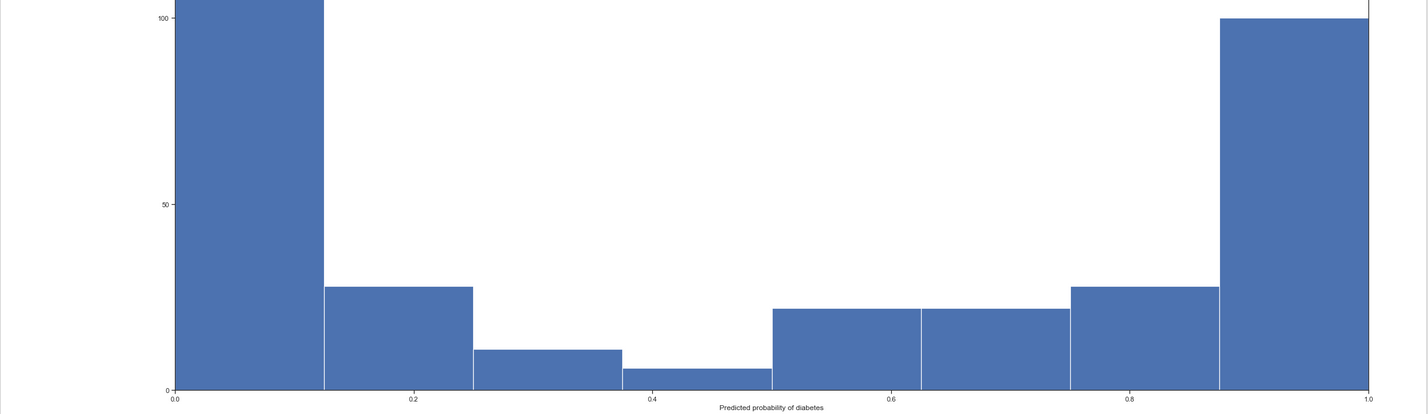


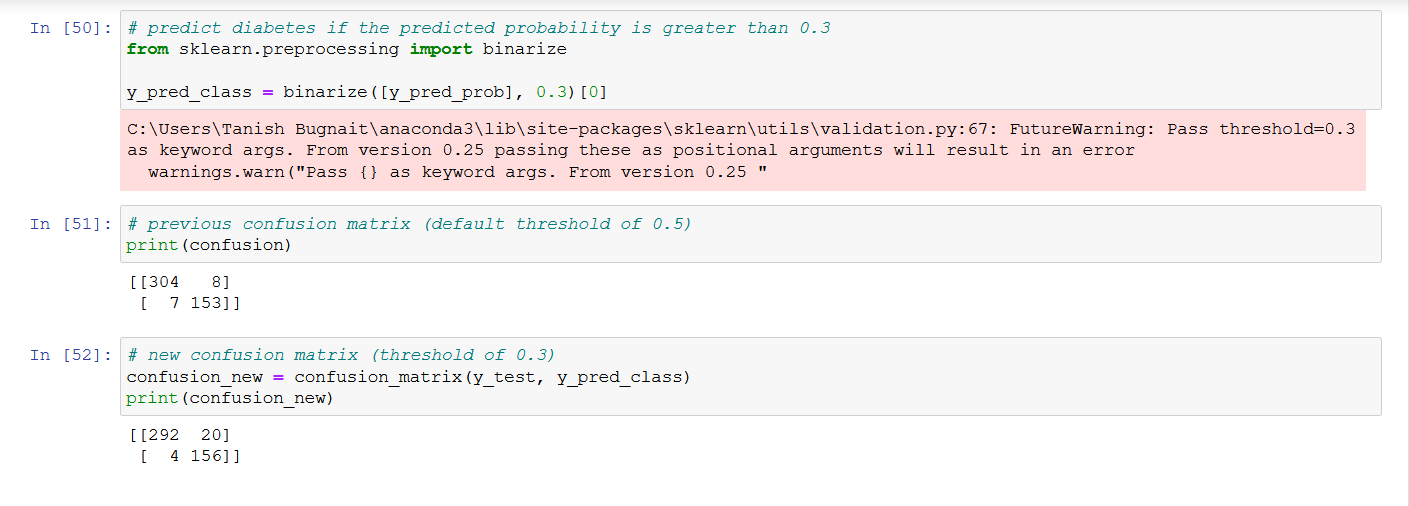




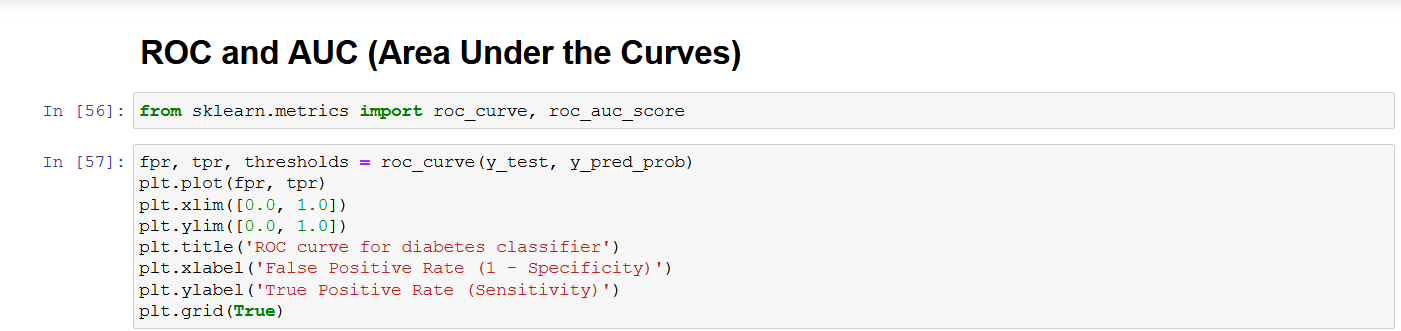




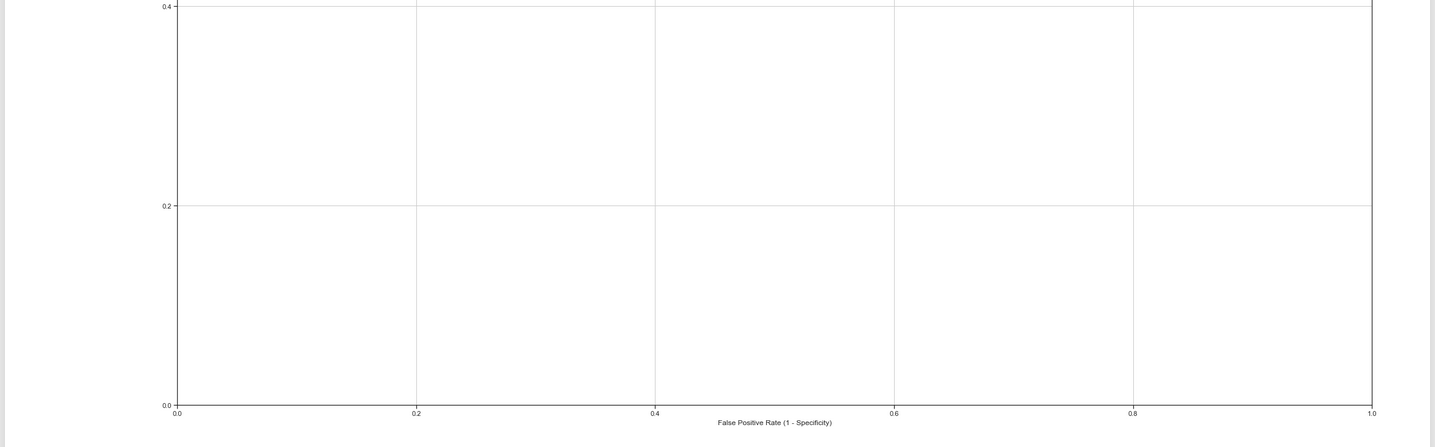




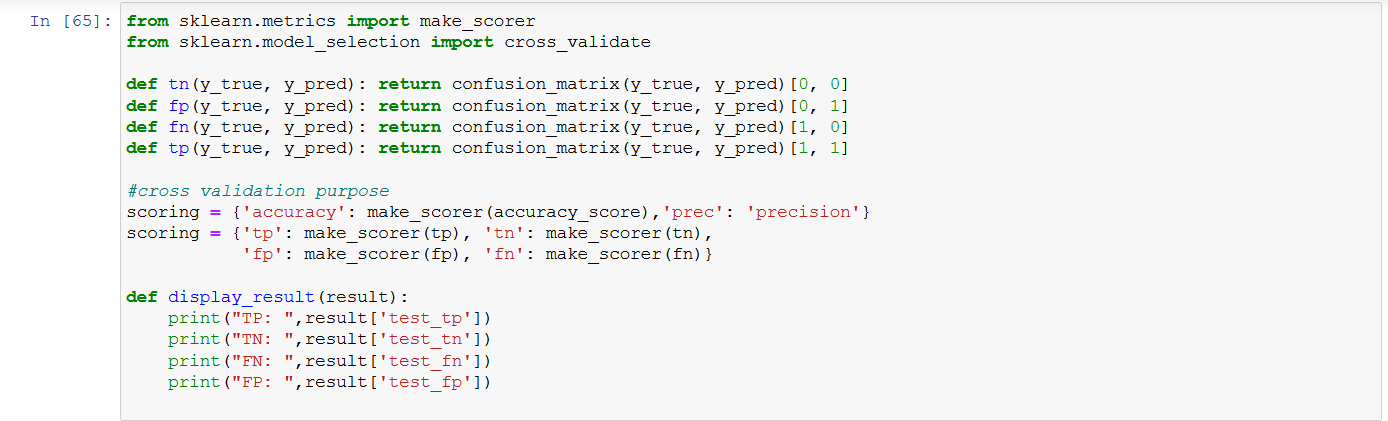


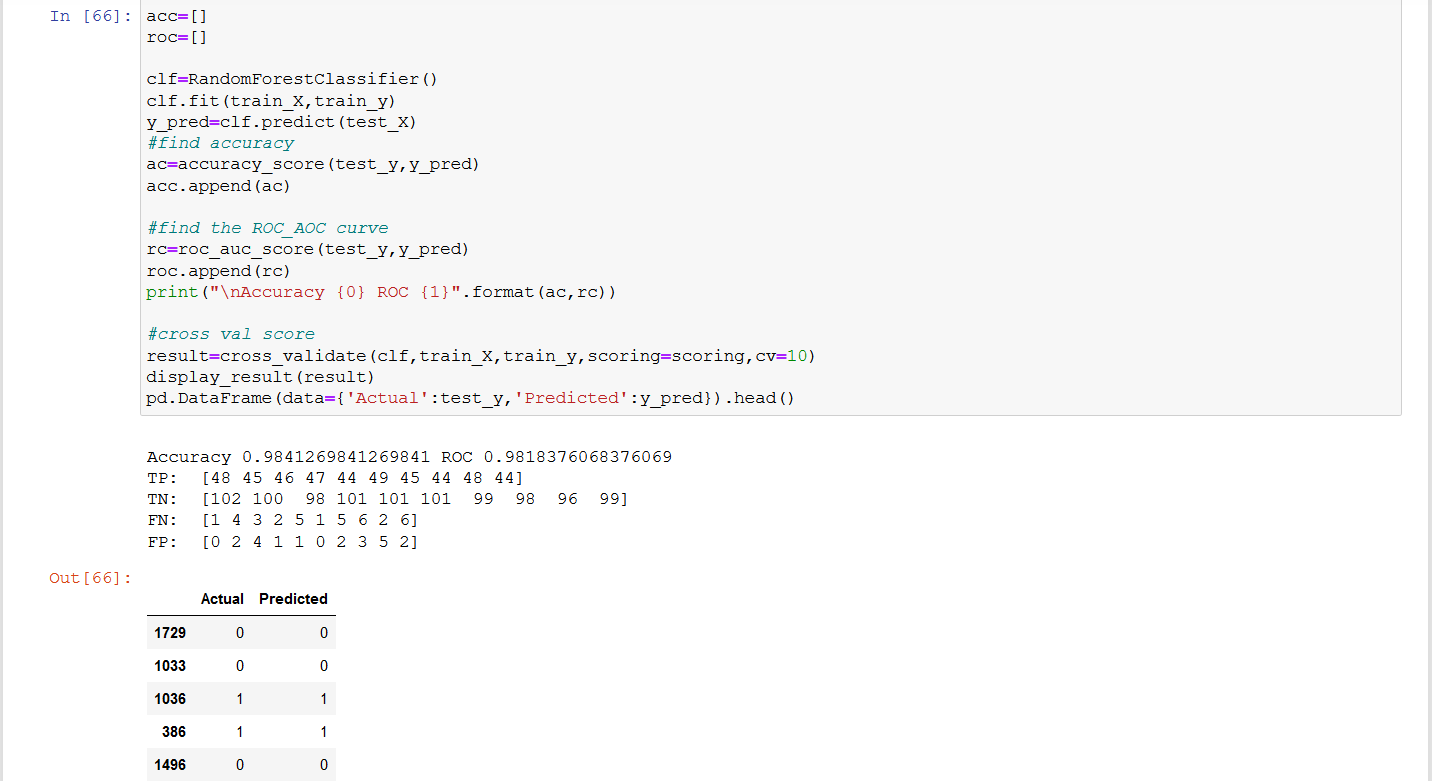




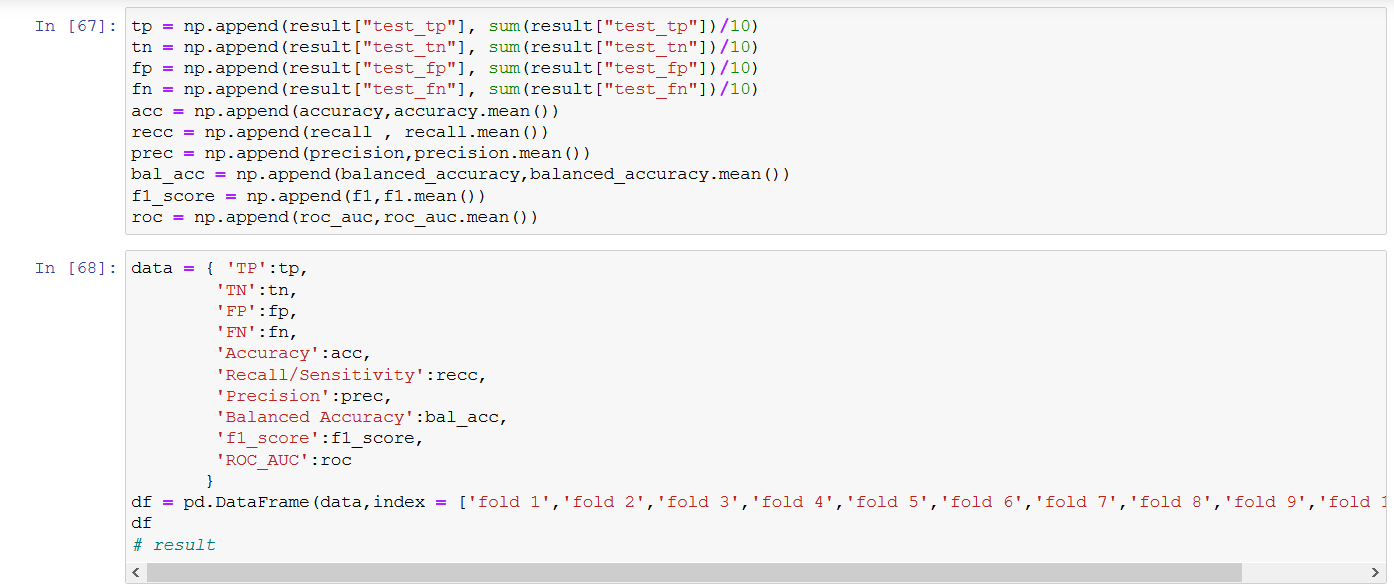


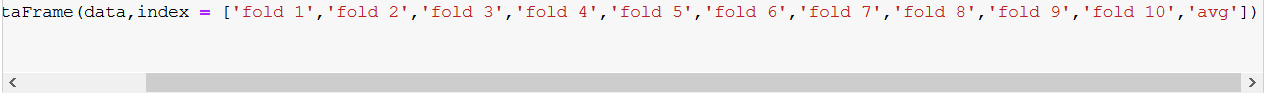


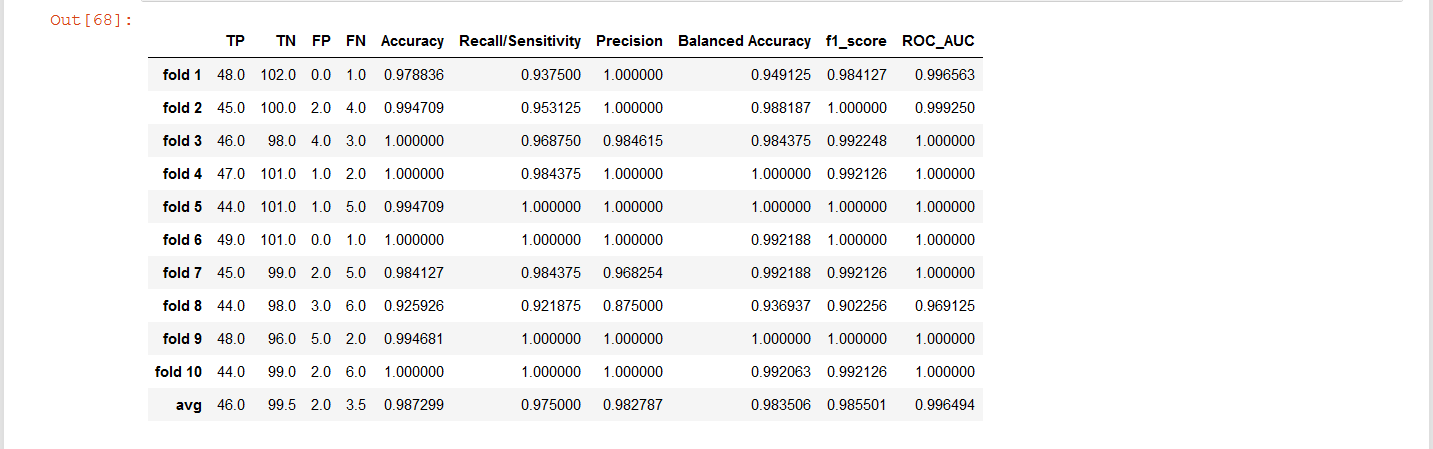




Now, we have a look at the various values in a tabular format to give a better clarity.



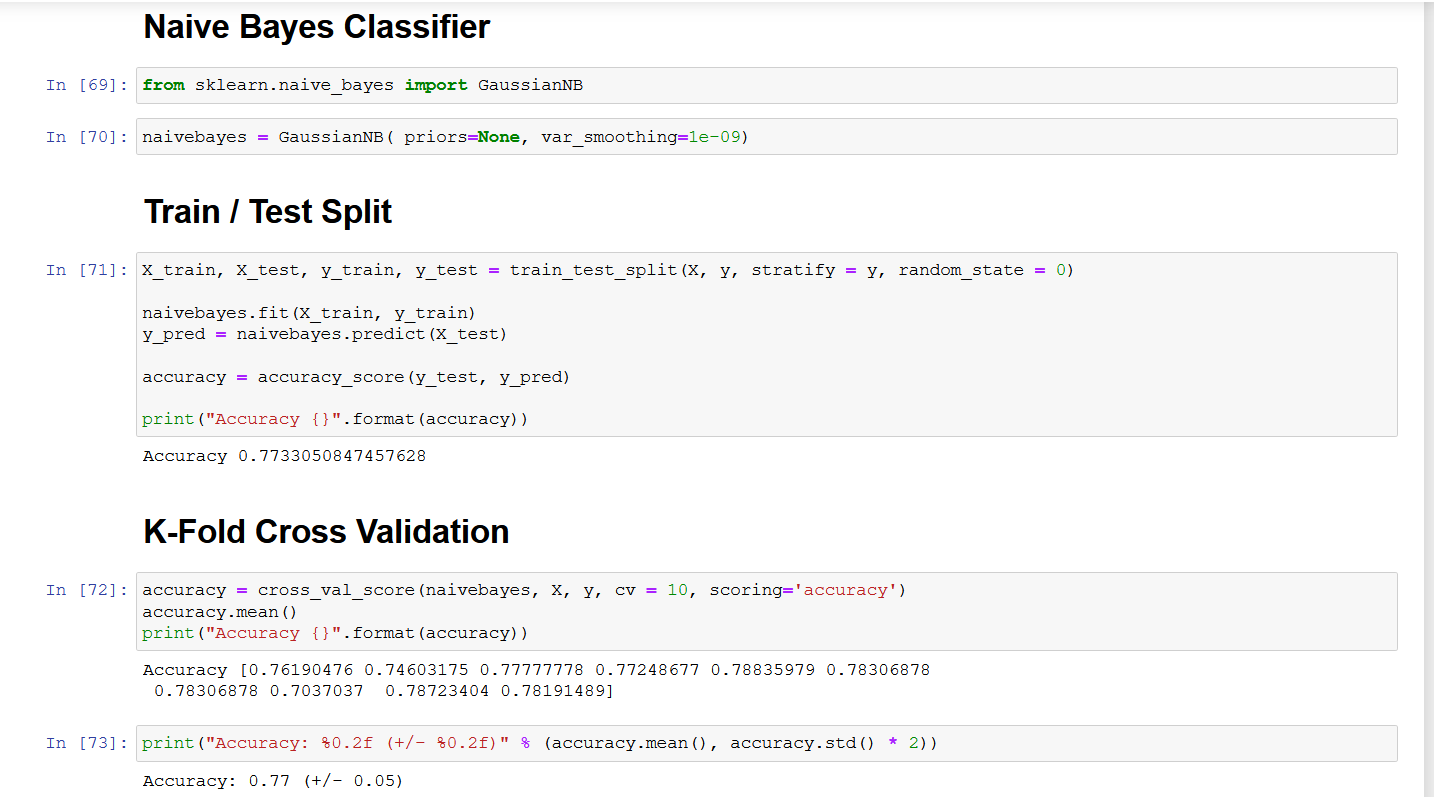




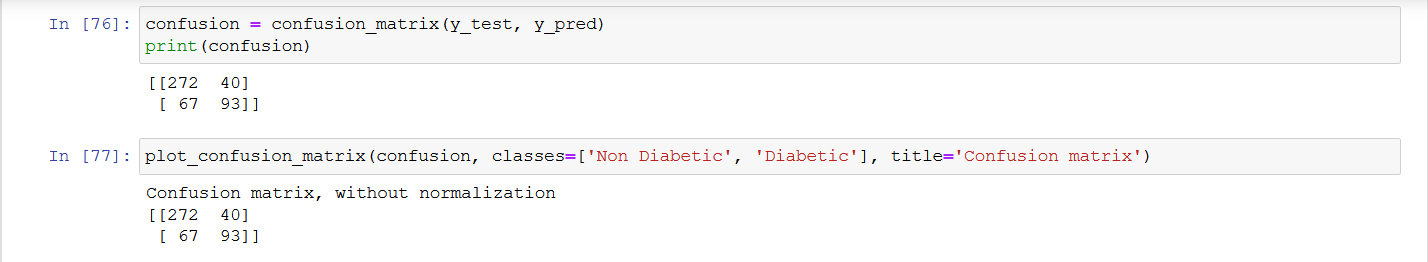
From the above screenshots we can see that I have calculated various metrics values for 10 folds and at the end I have also calculated the average as it was desired for the project.

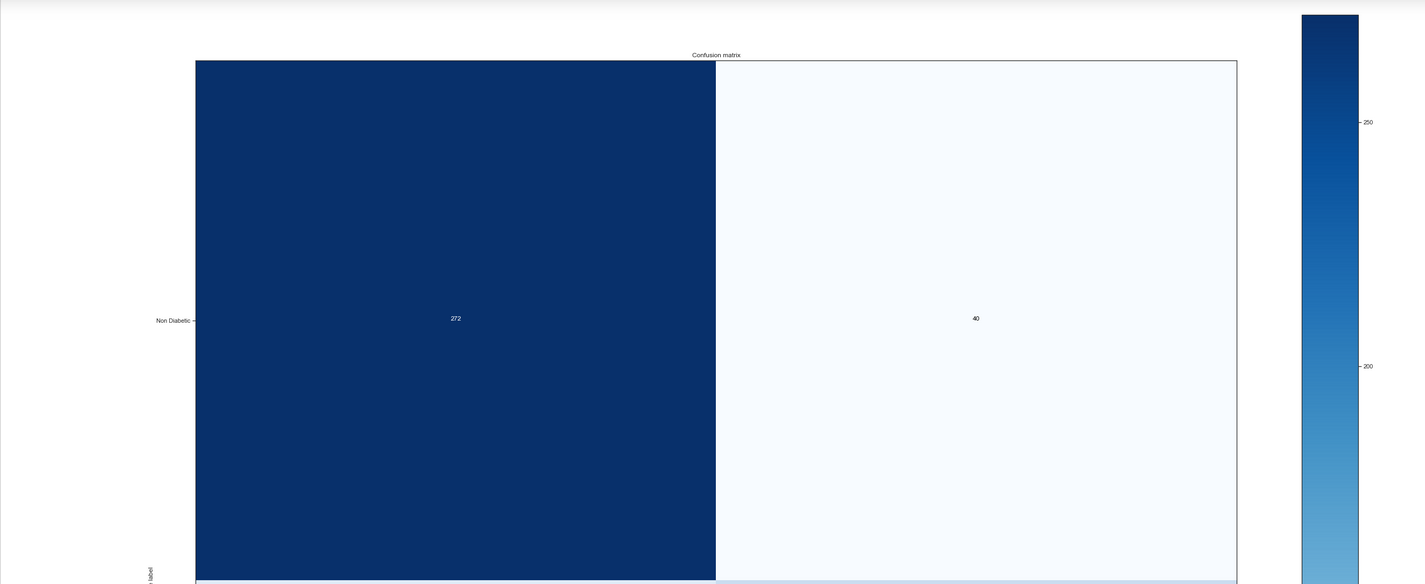
Now, that we have seen Random Forest classifier we will shift our focus to Naïve Bayes Classifier.

1. **Naïve Bayes**

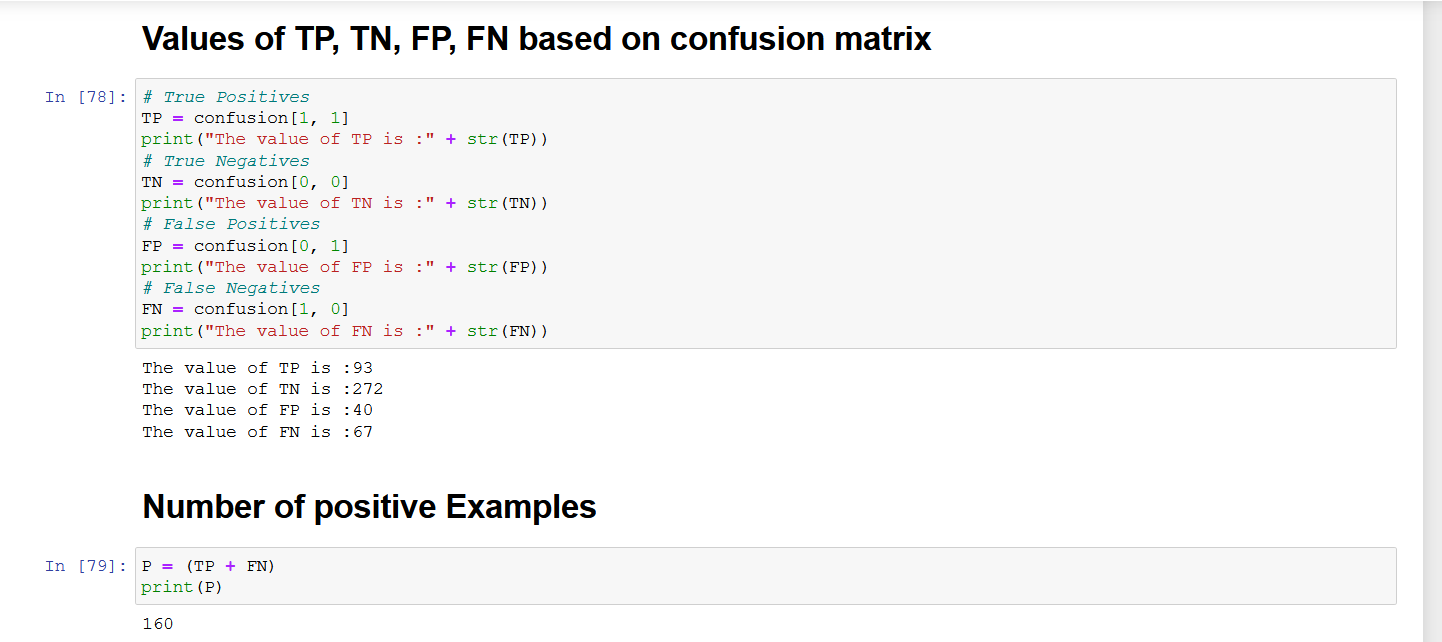


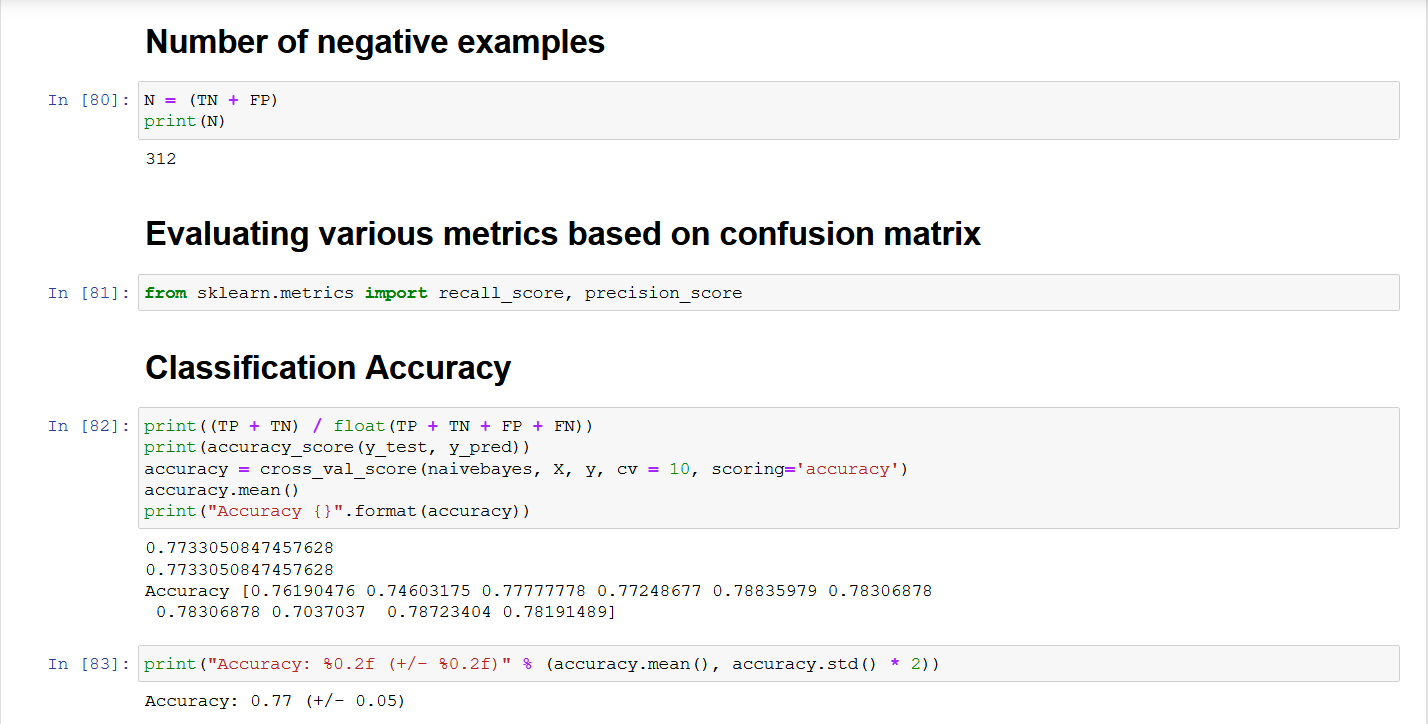


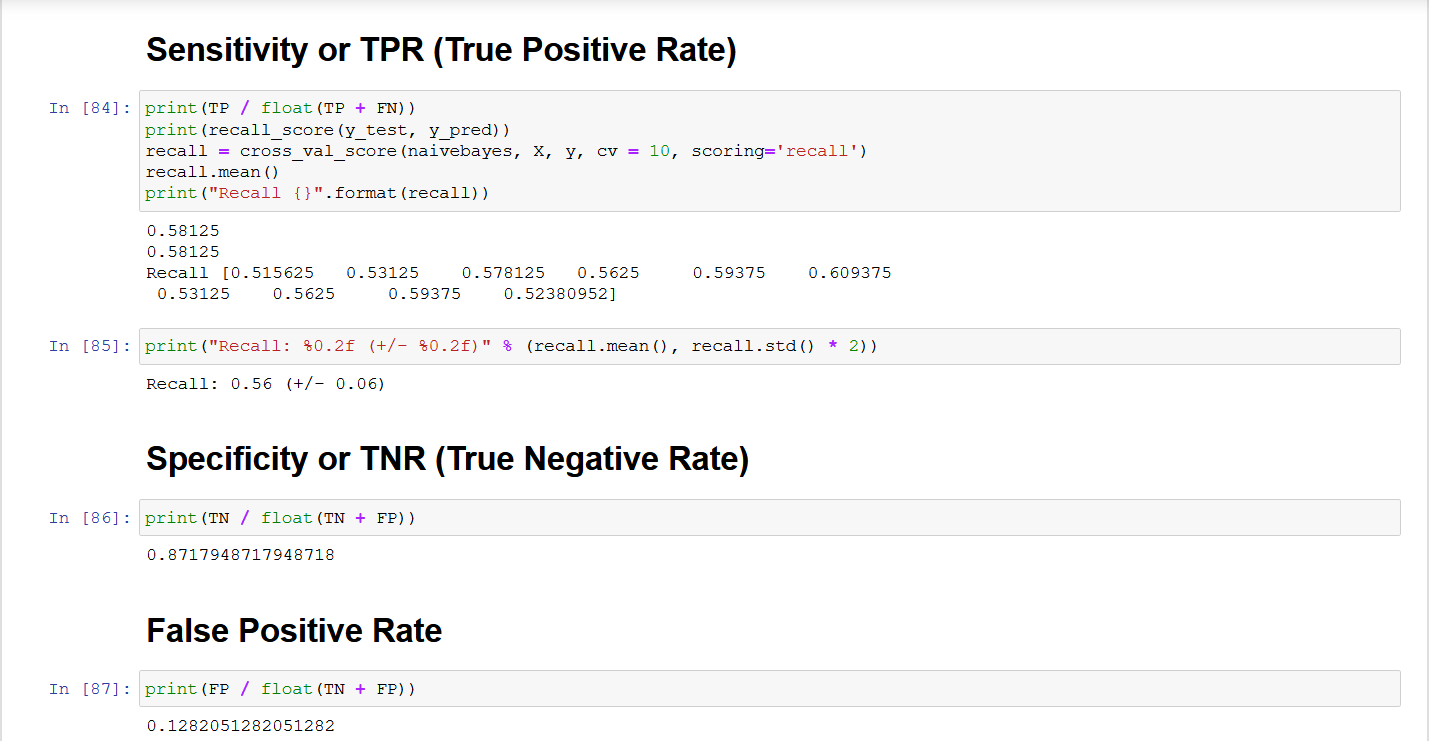






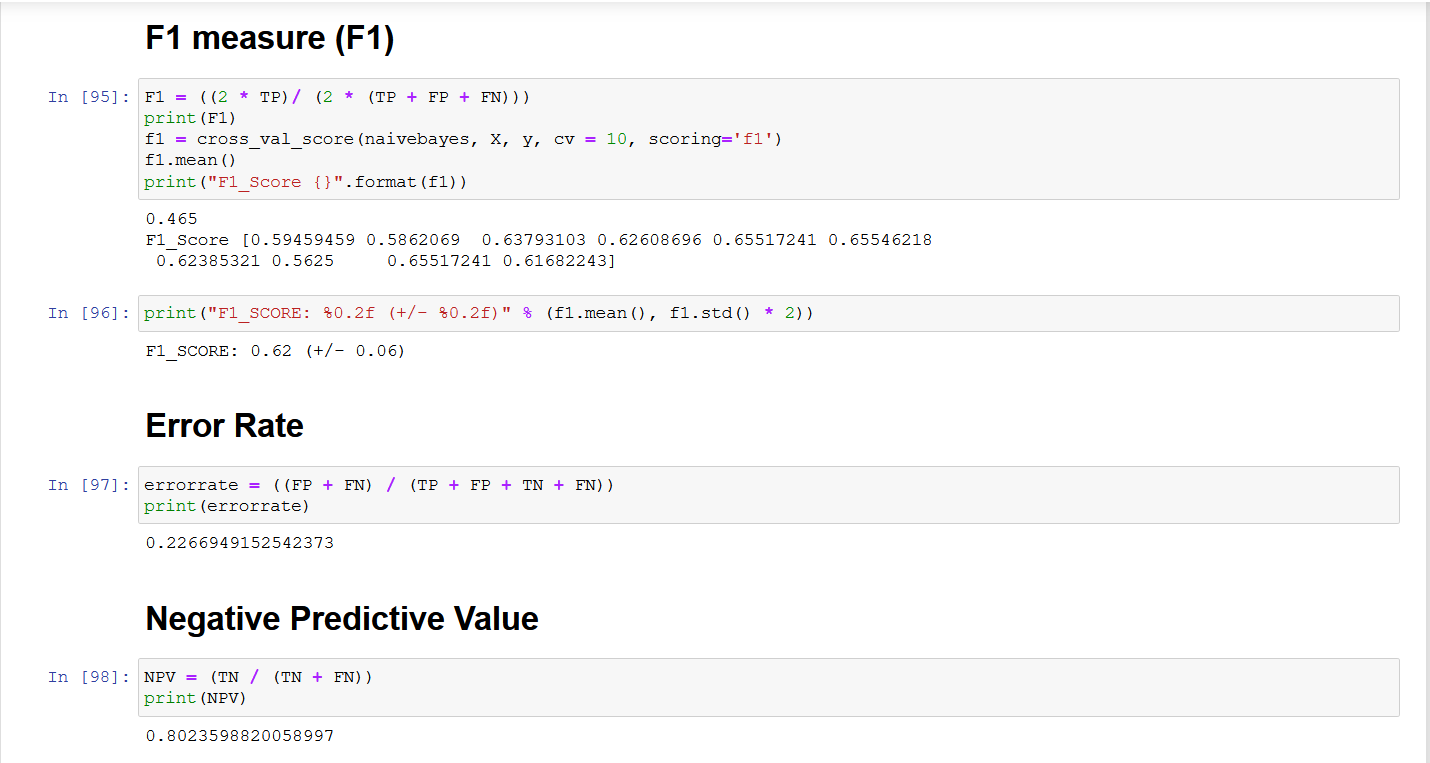


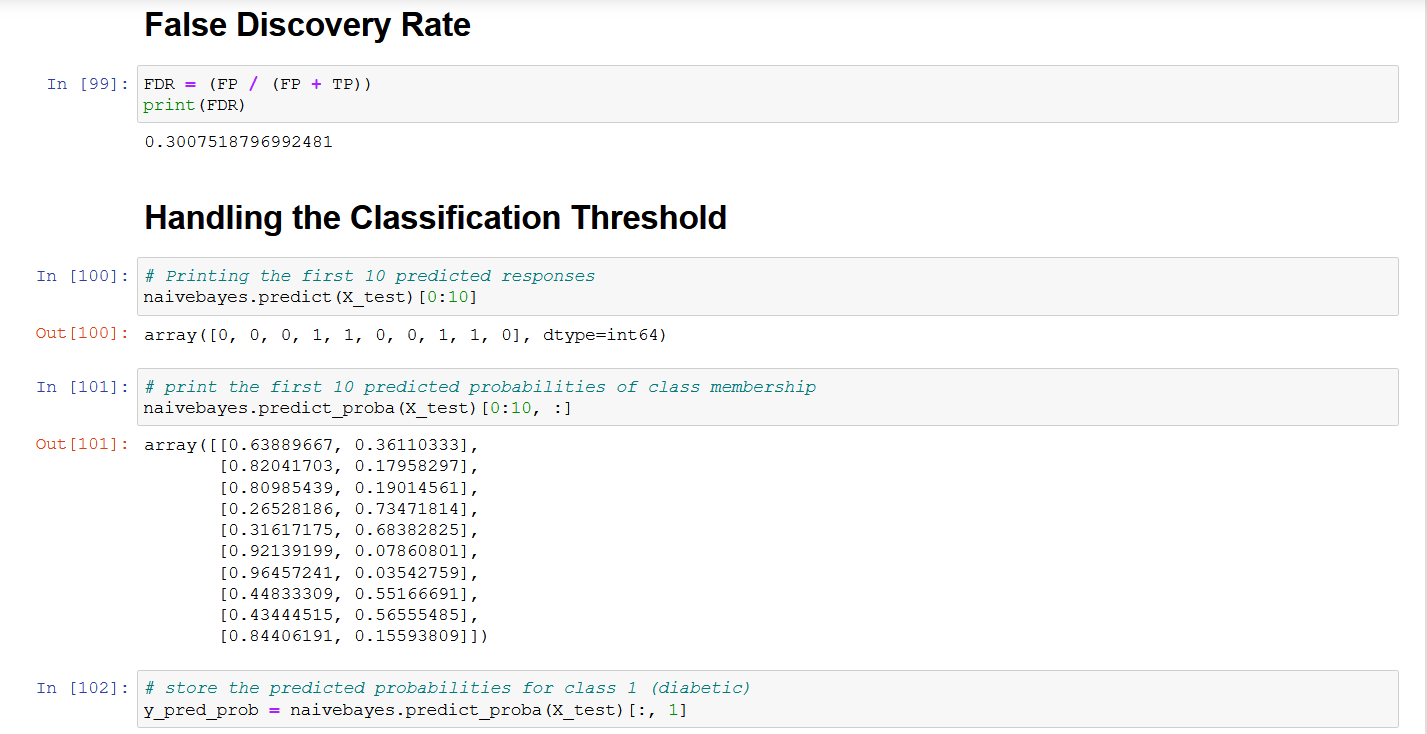




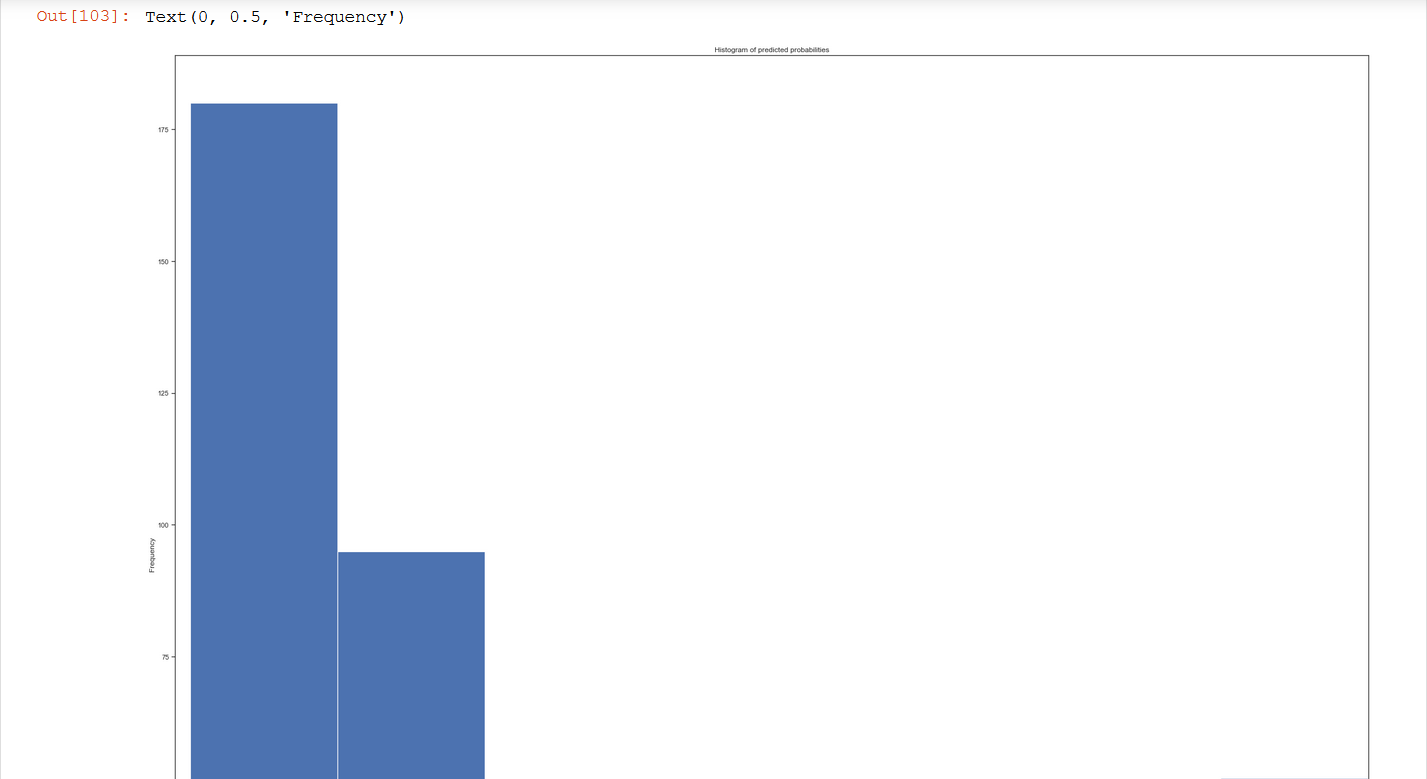


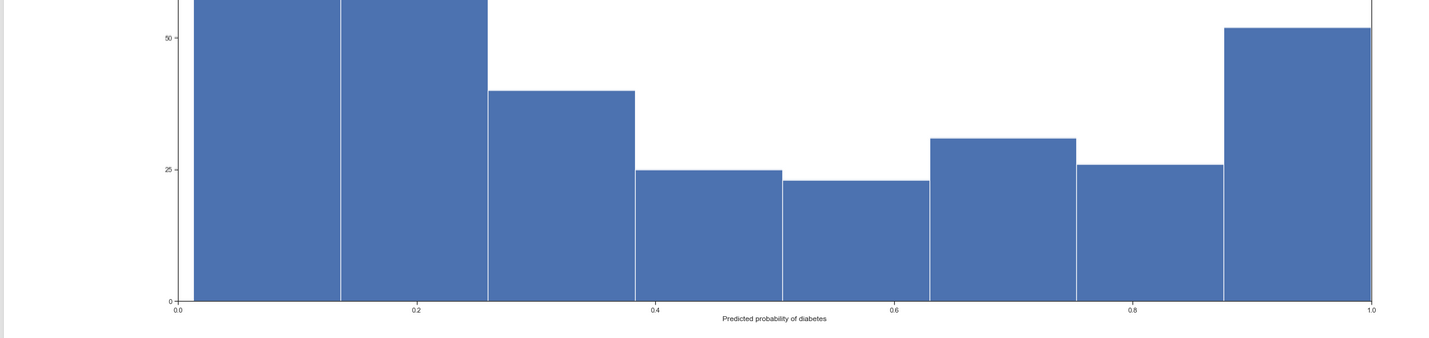


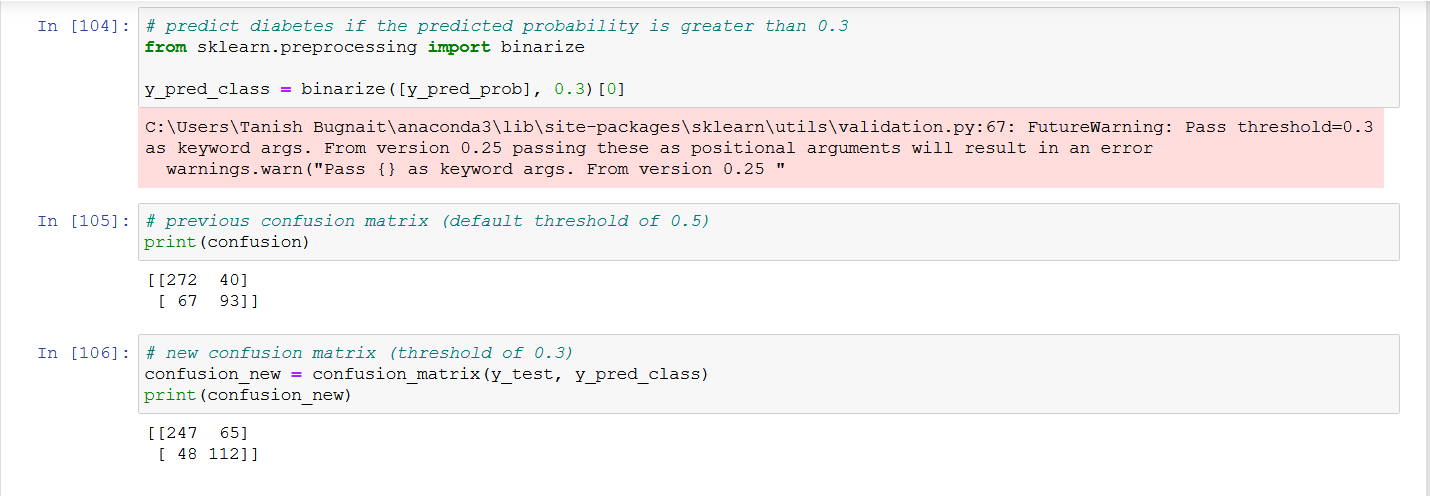


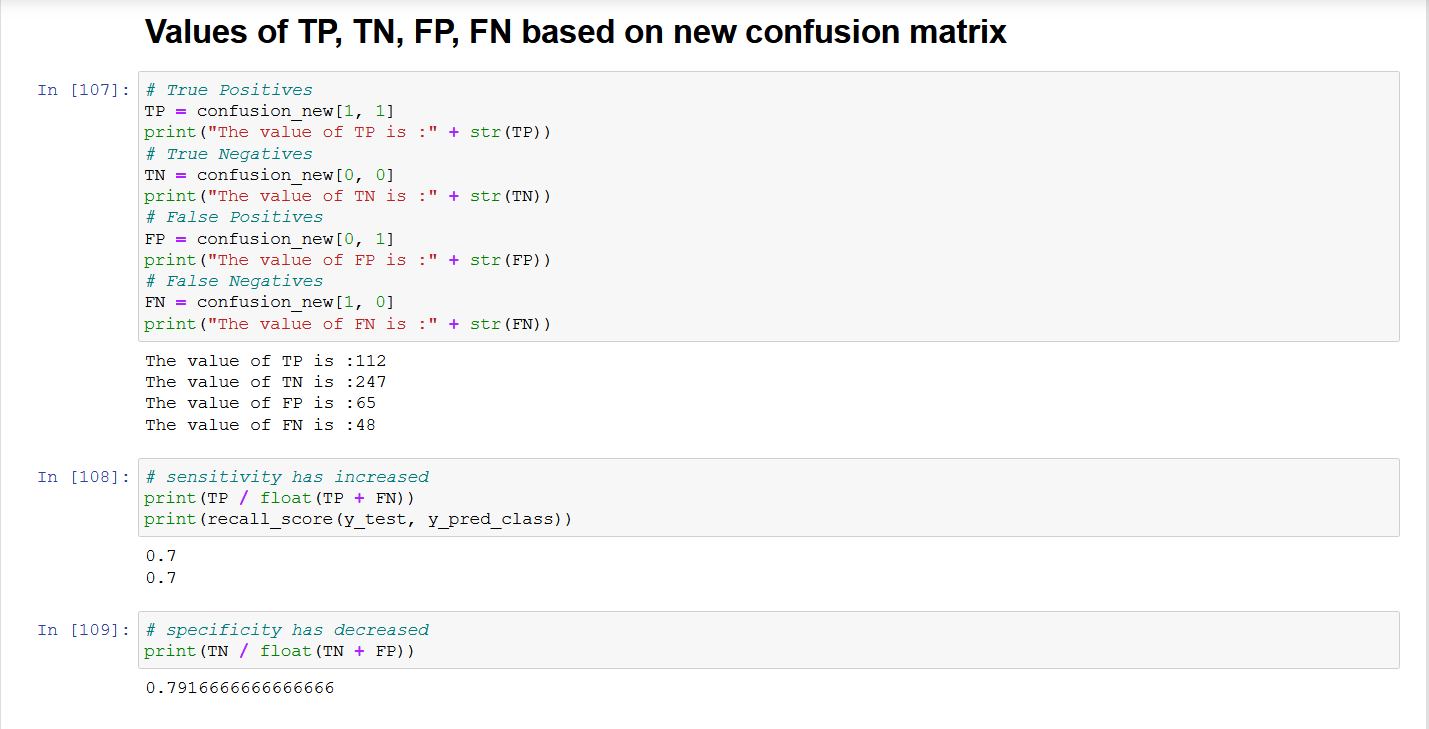


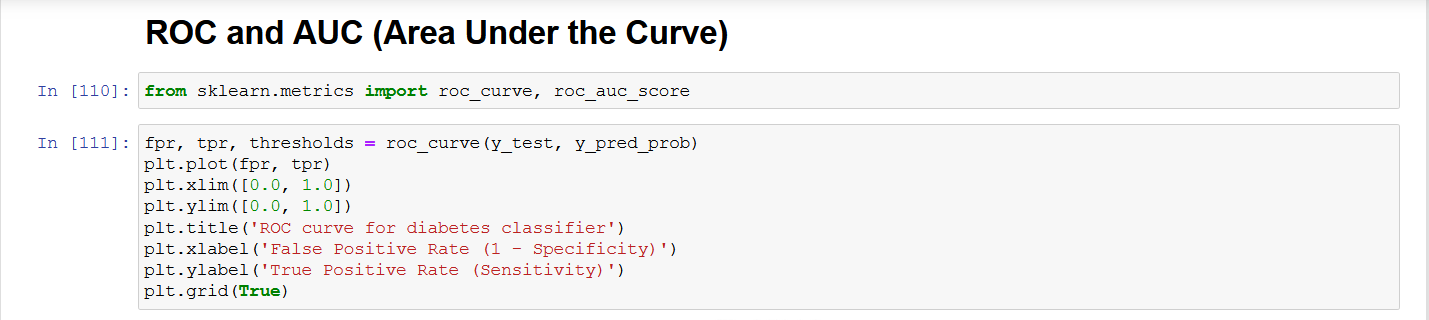


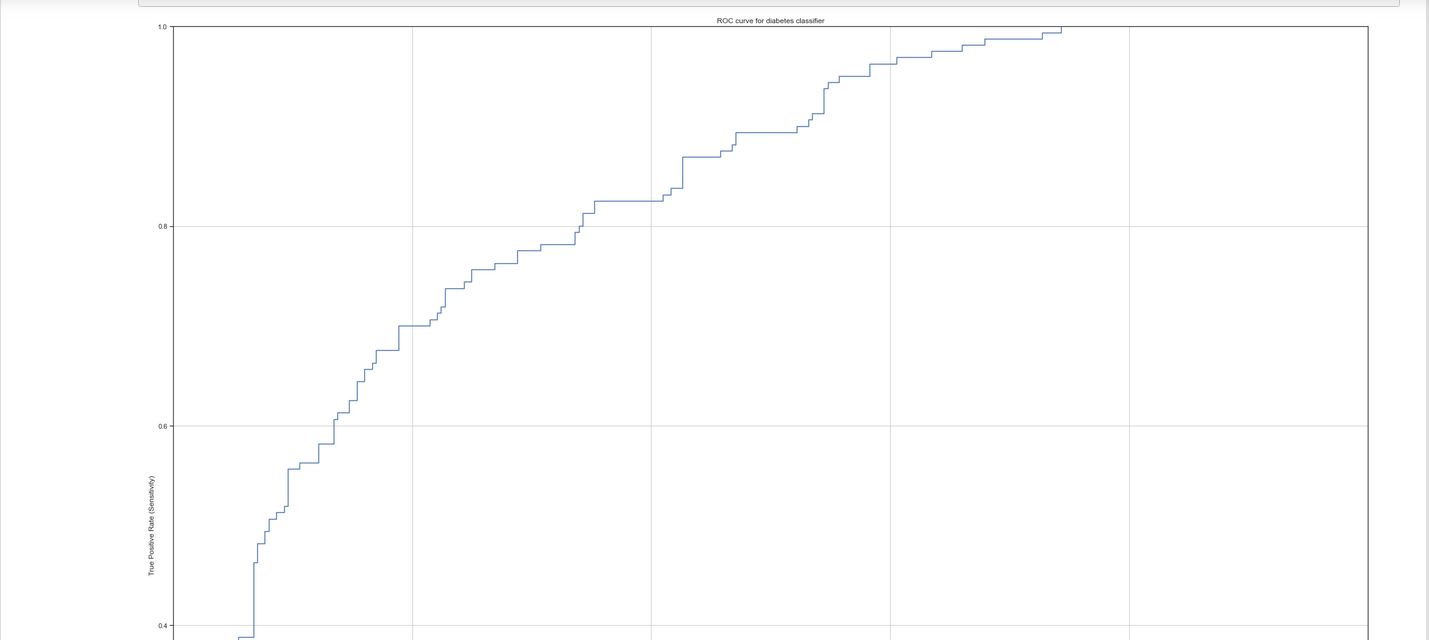


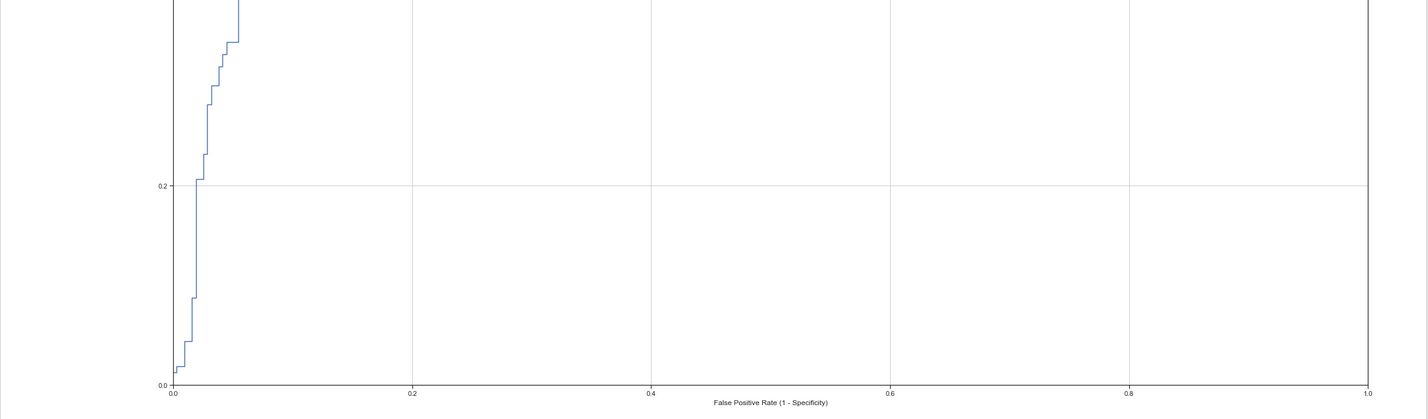


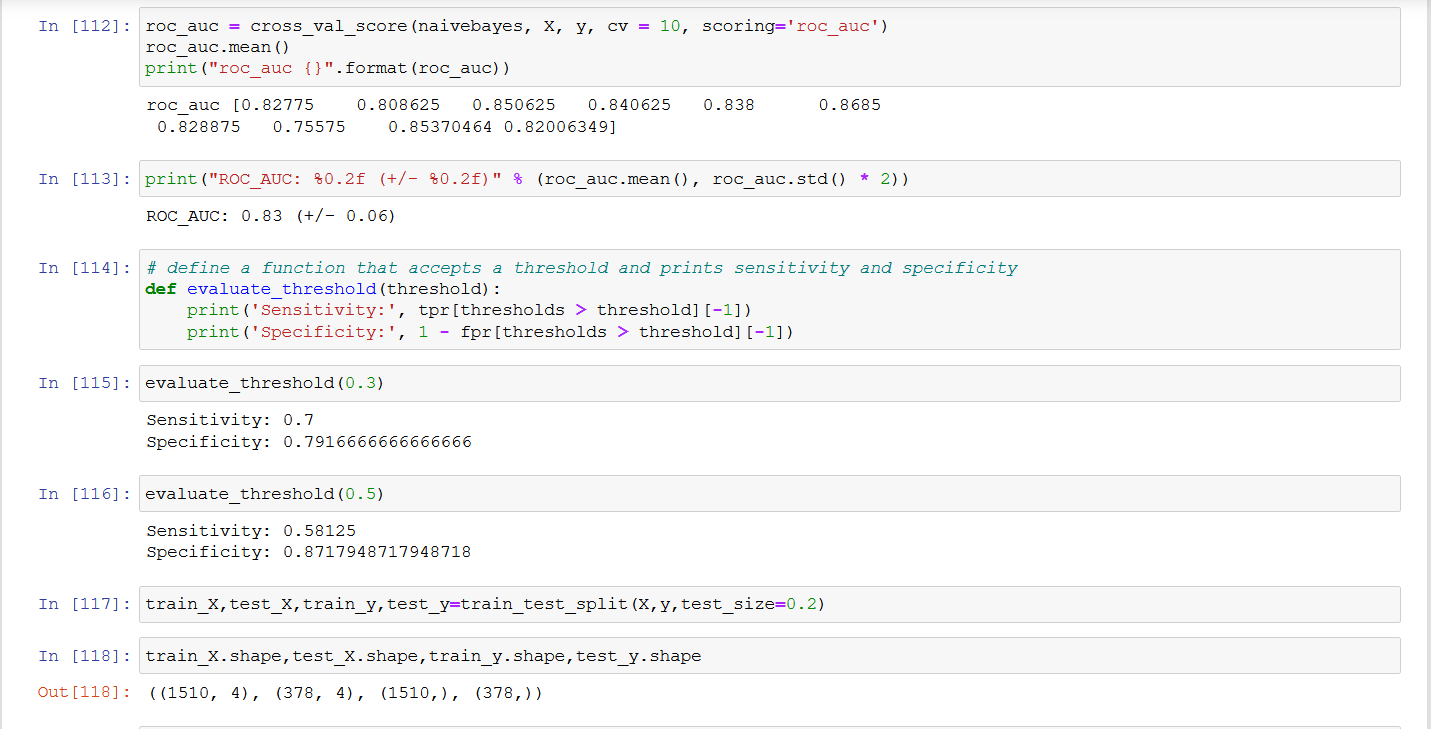


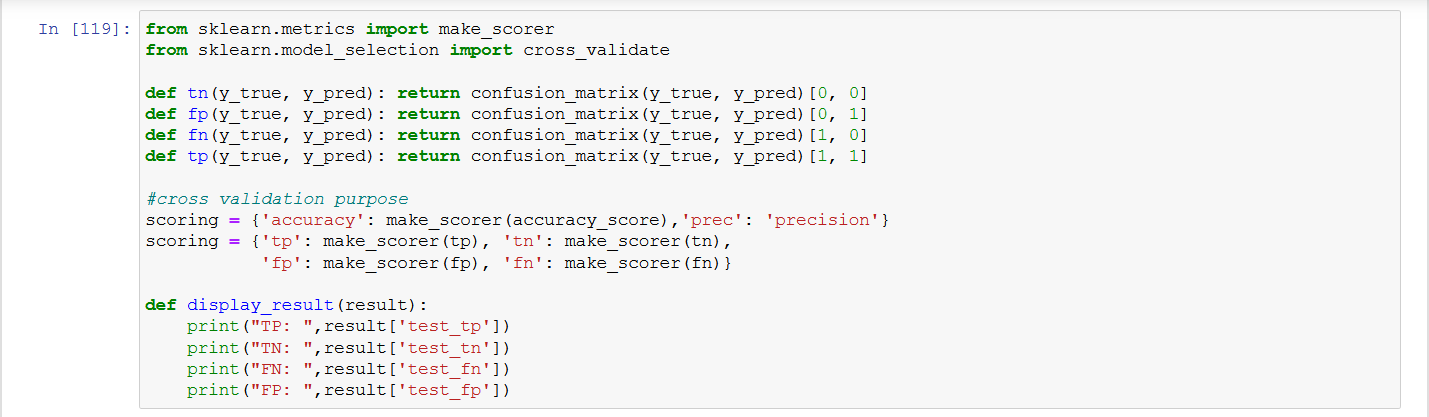




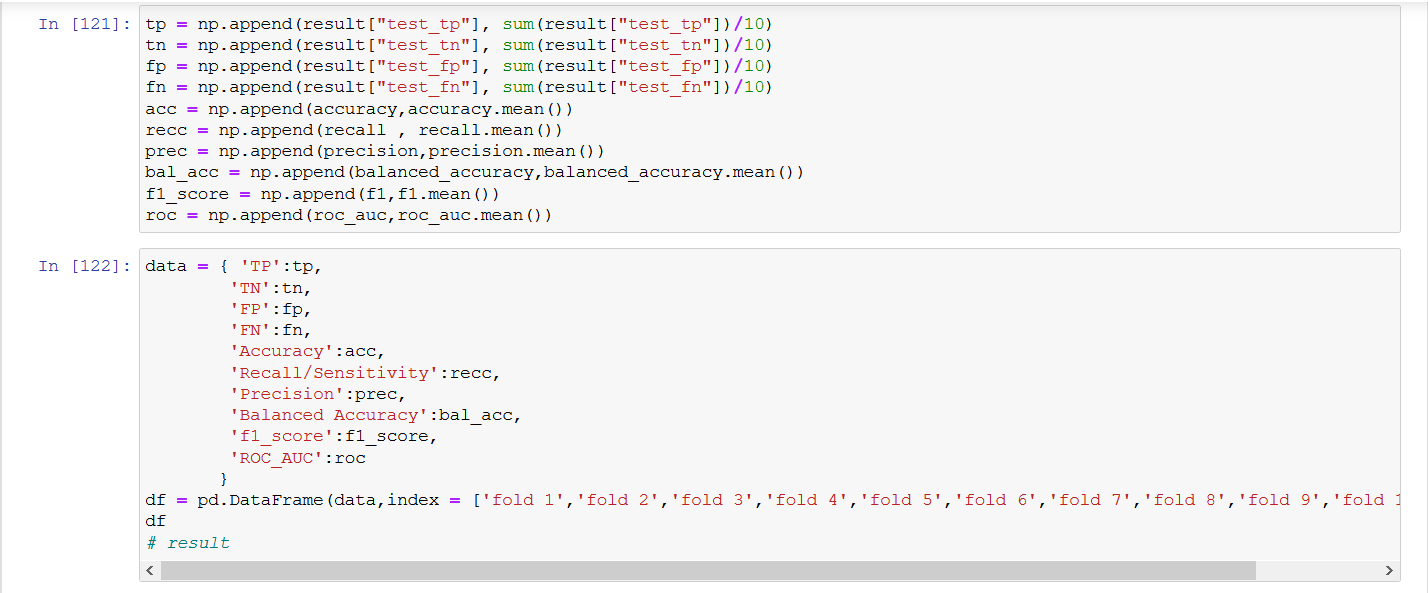


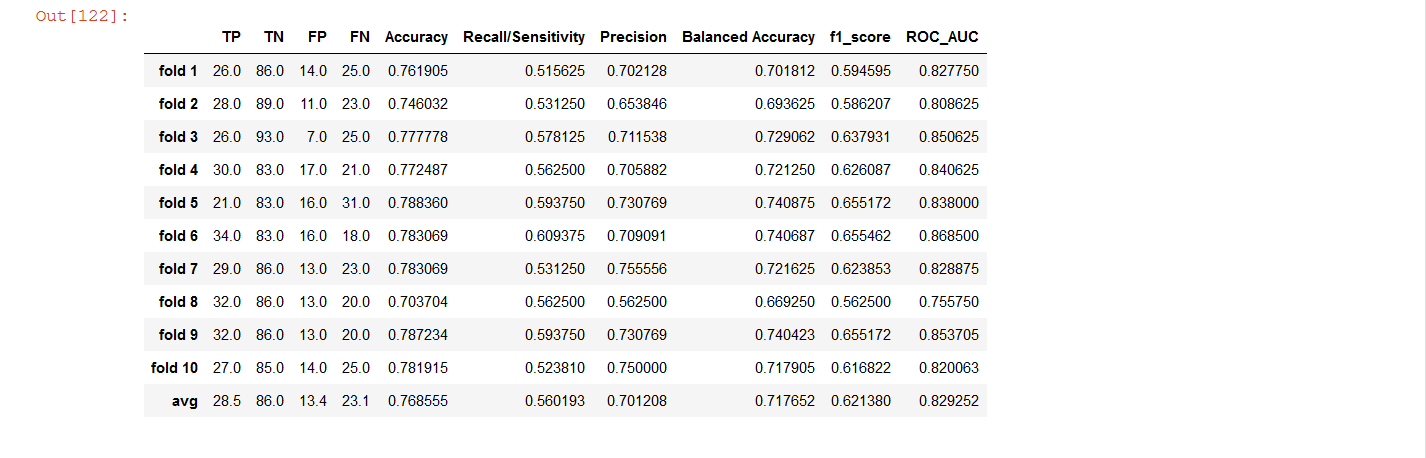




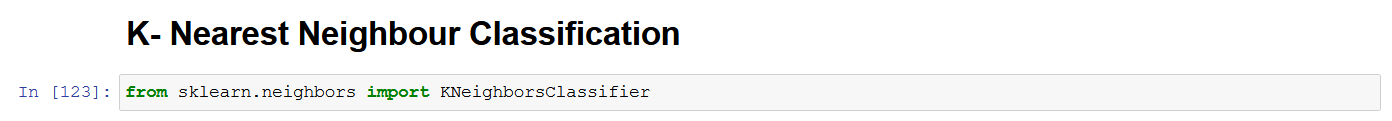








1. **KNN Classifier**

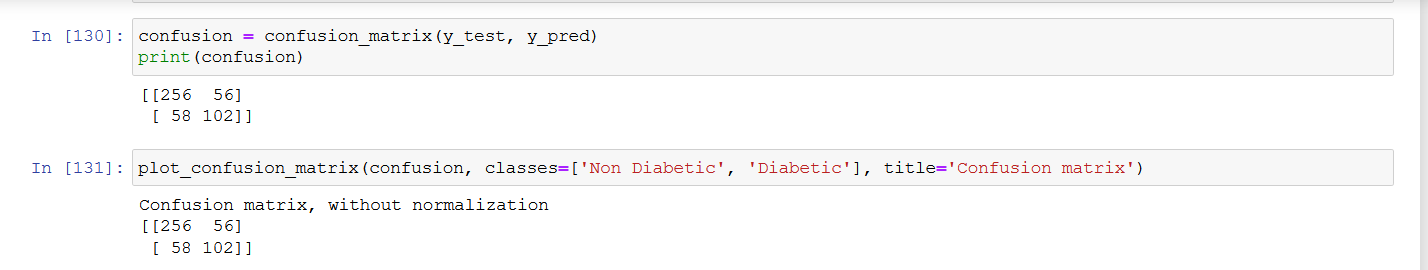






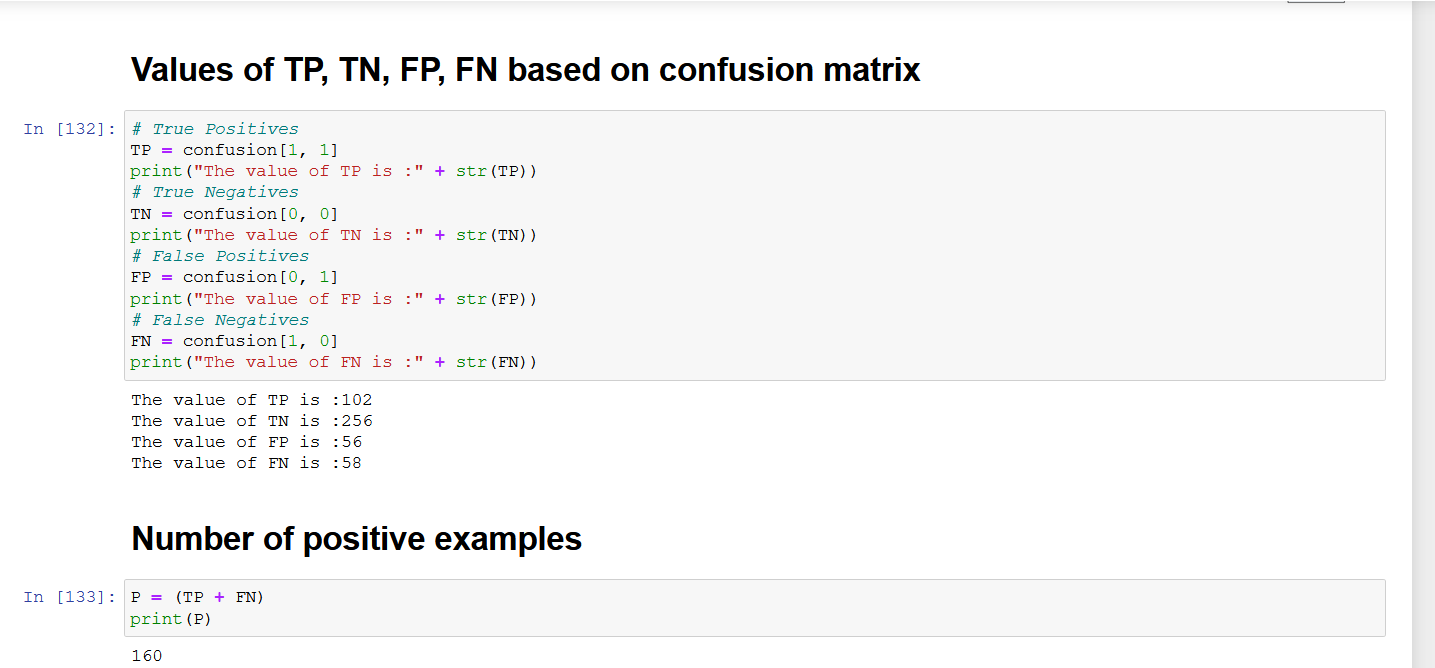


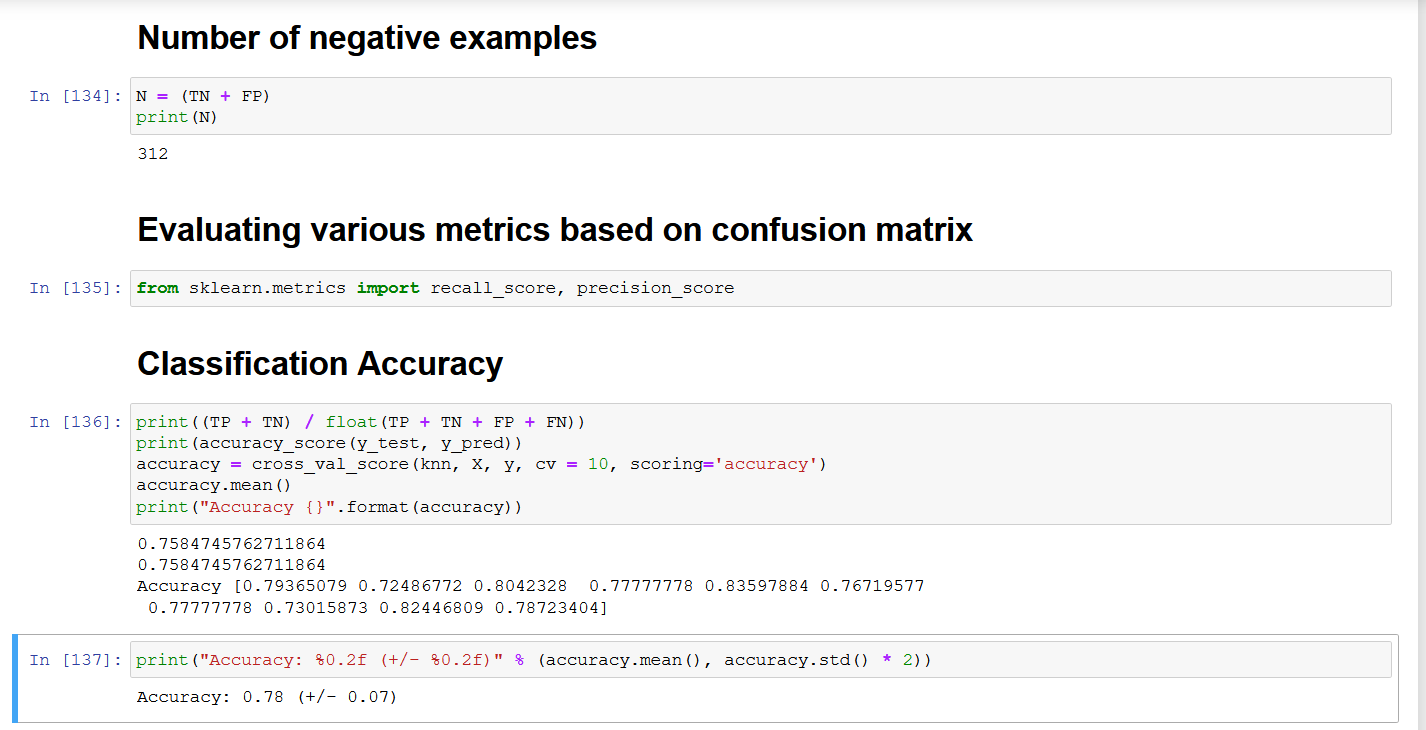






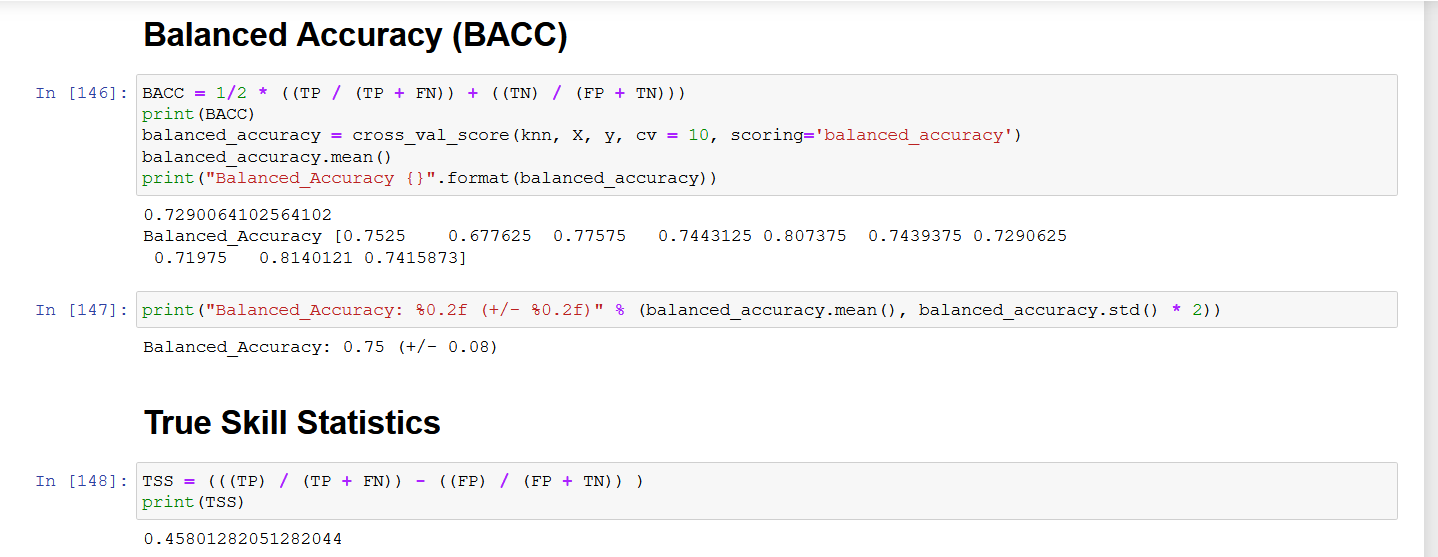




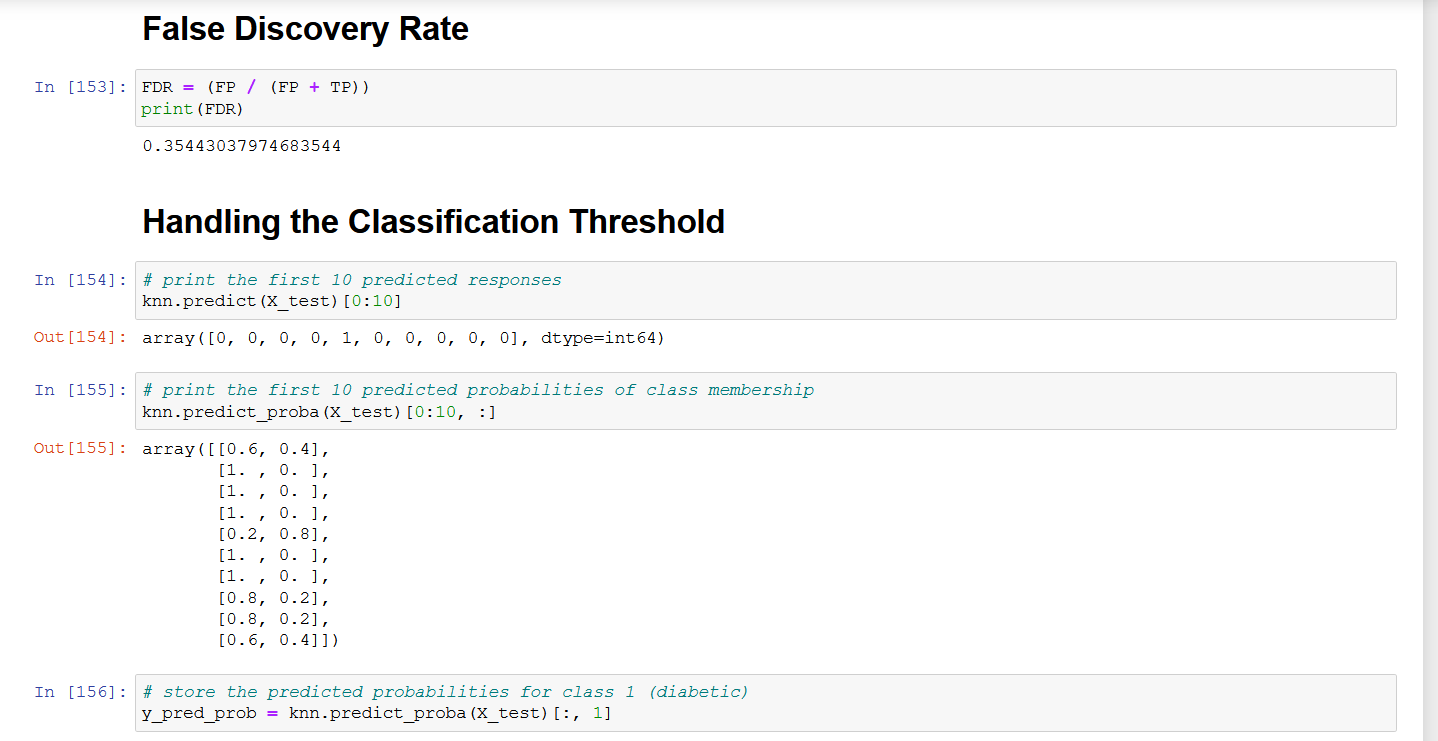


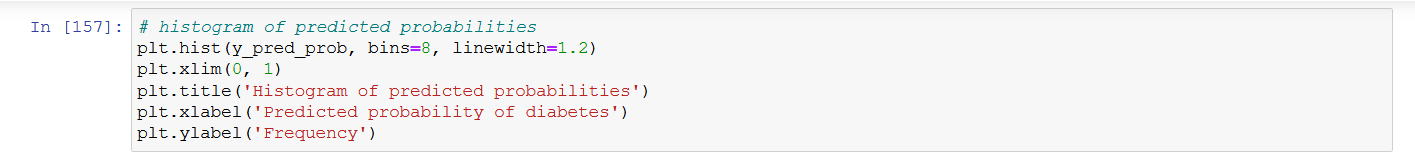


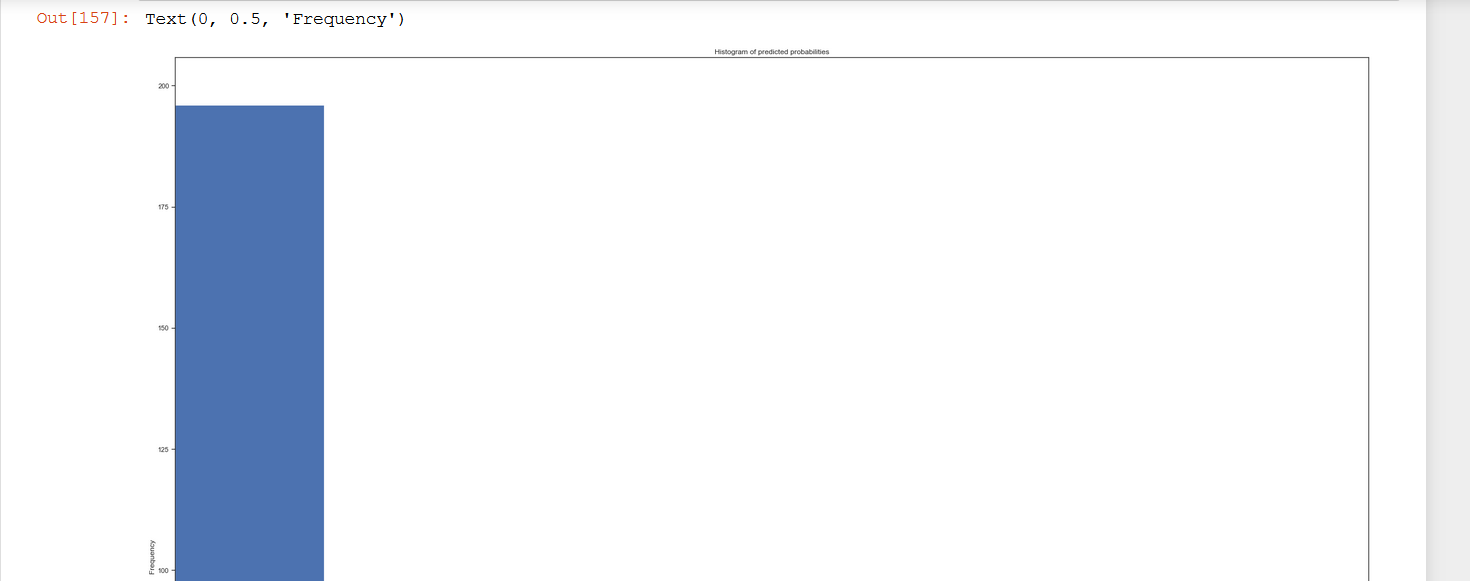


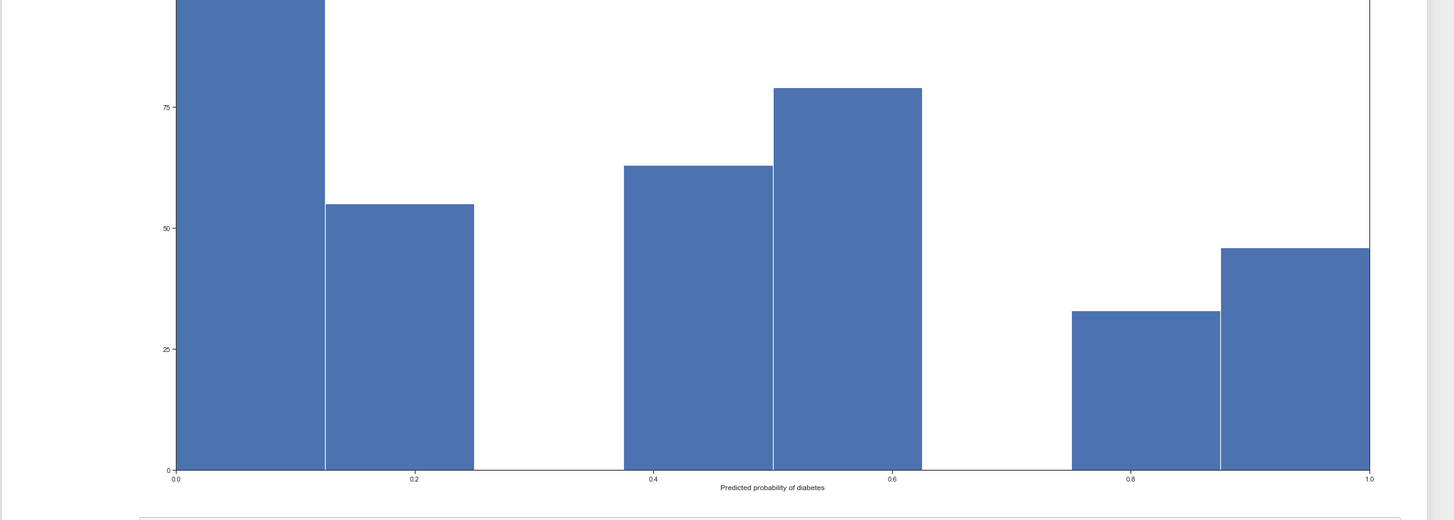


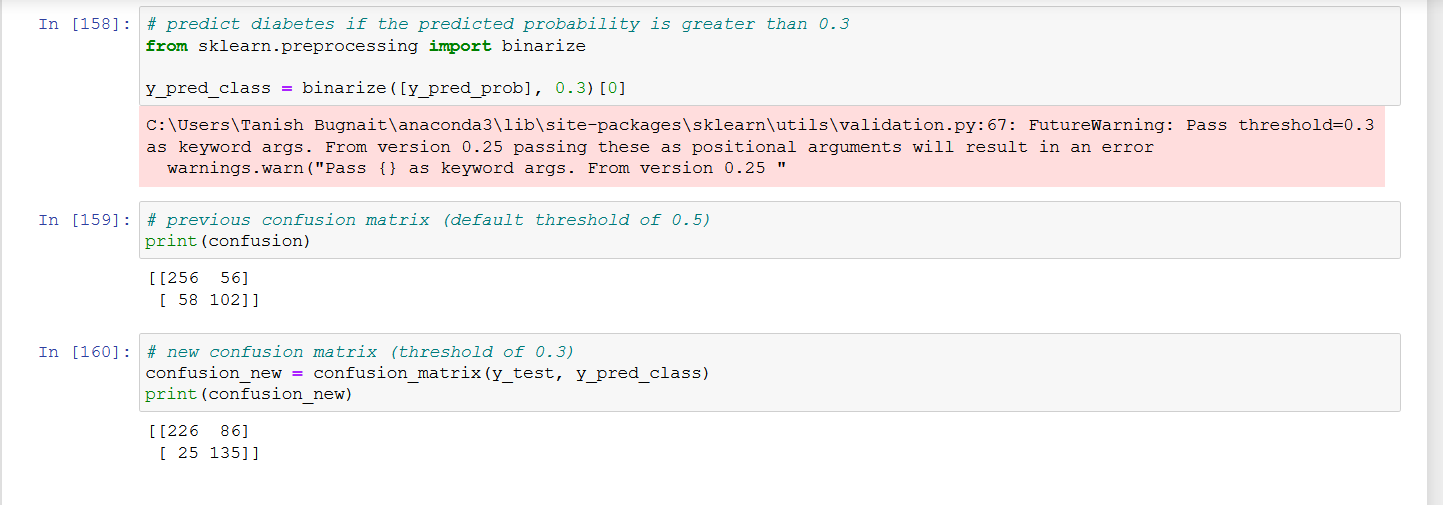


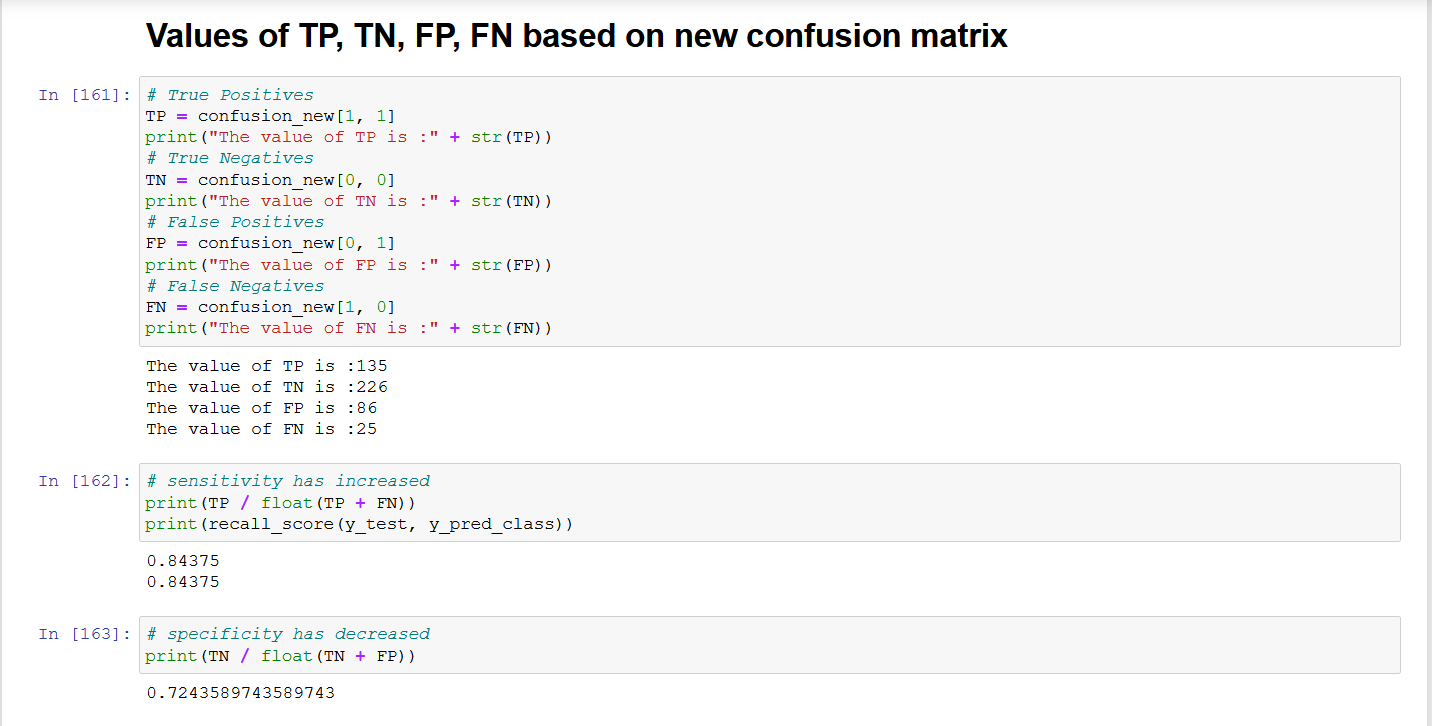


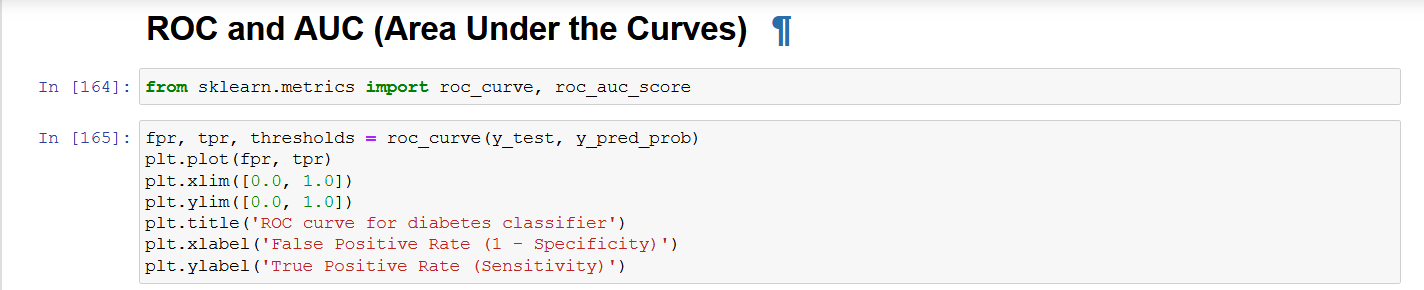


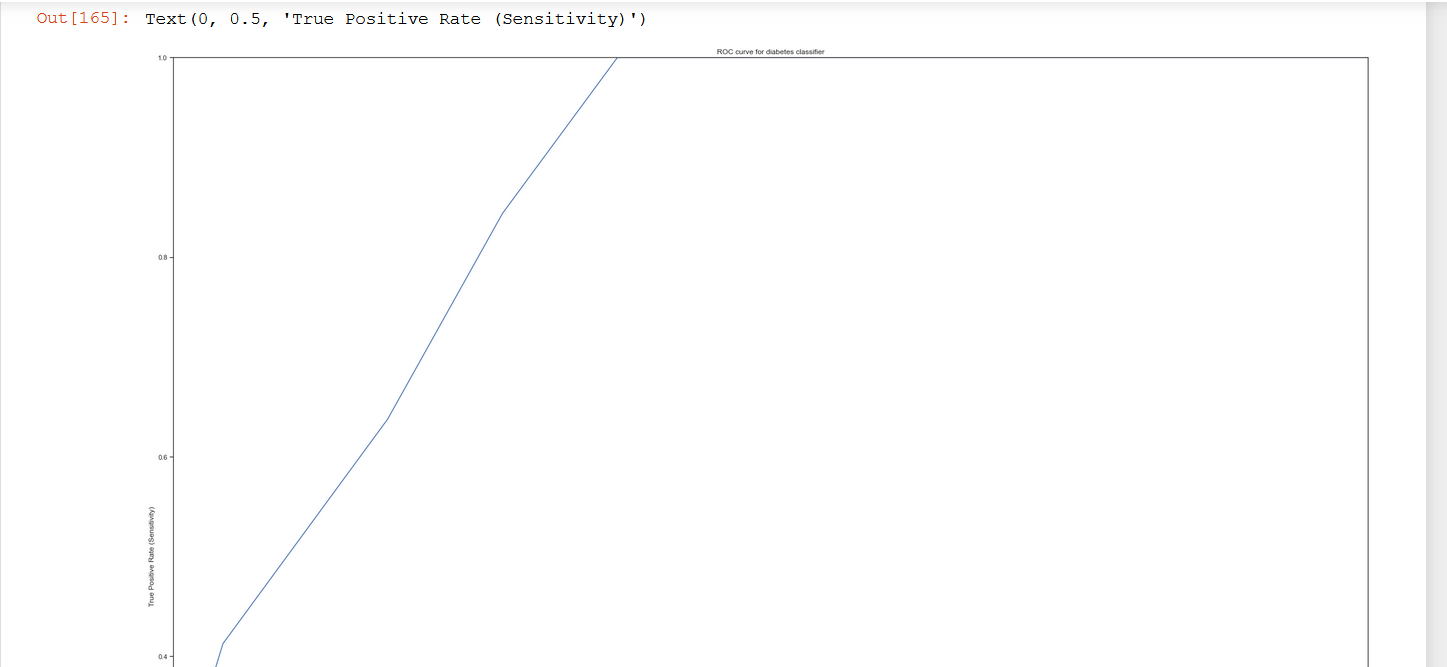


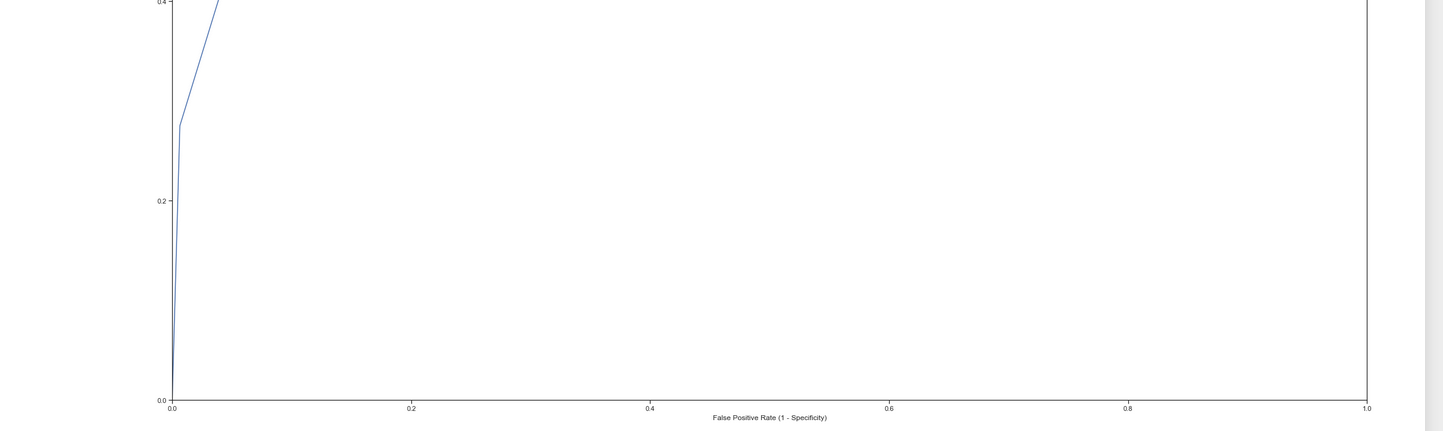


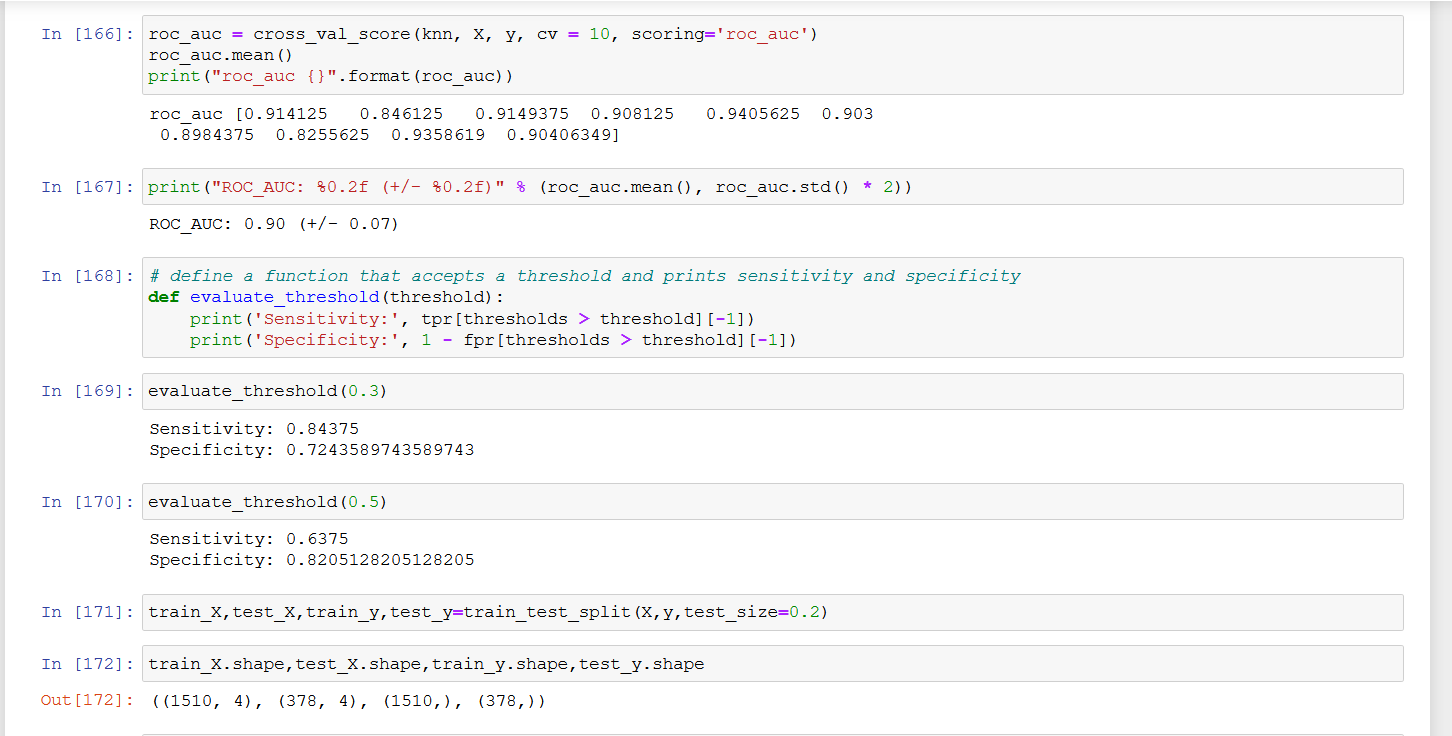


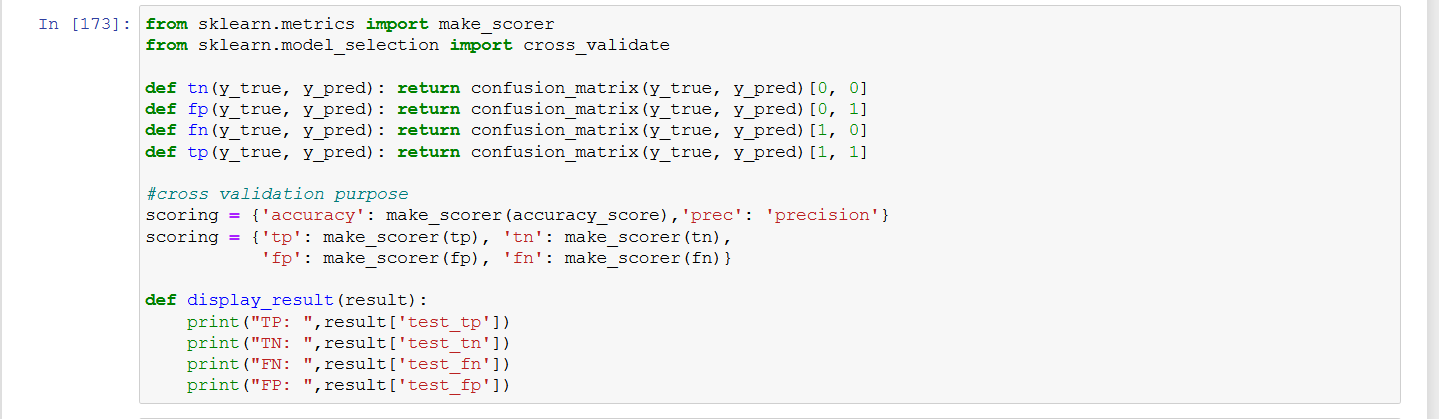


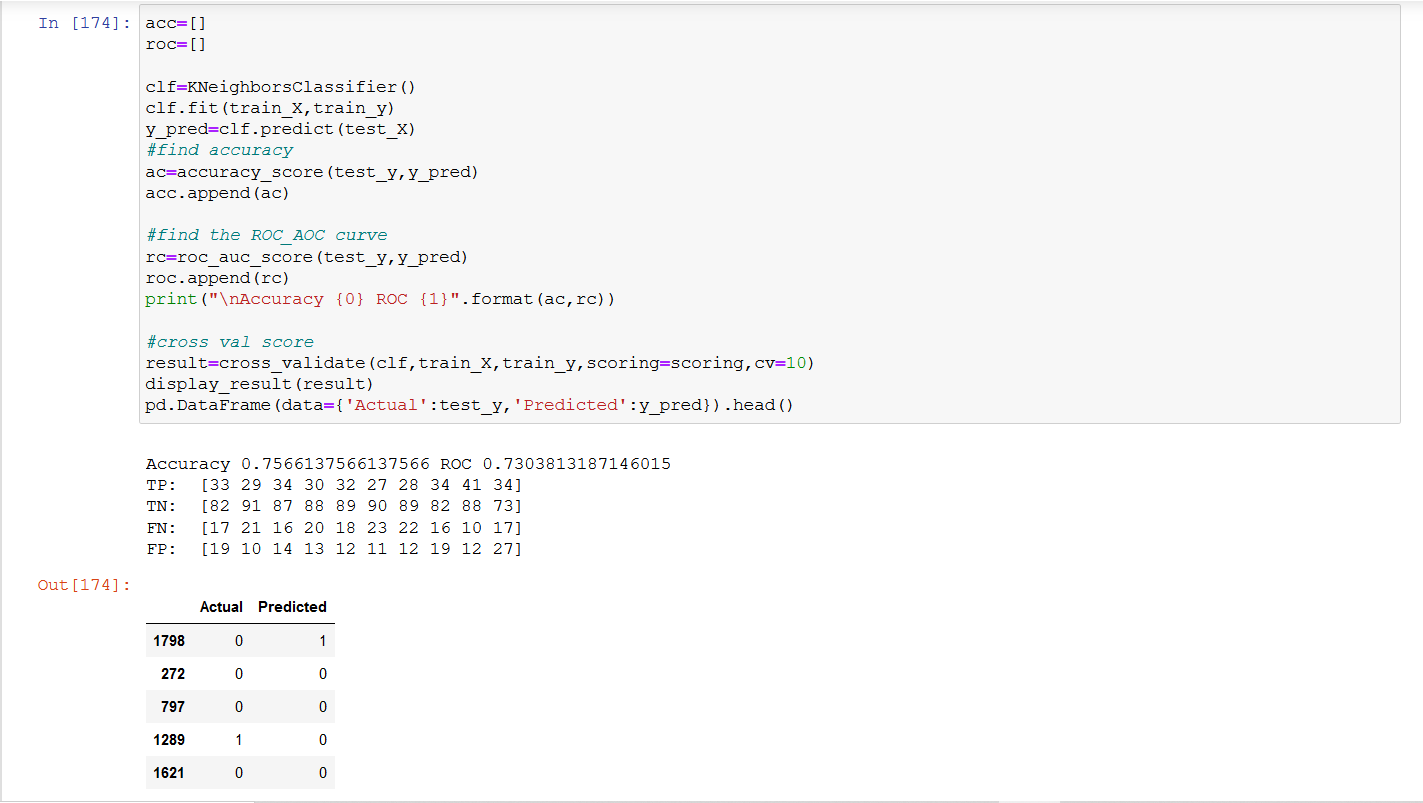


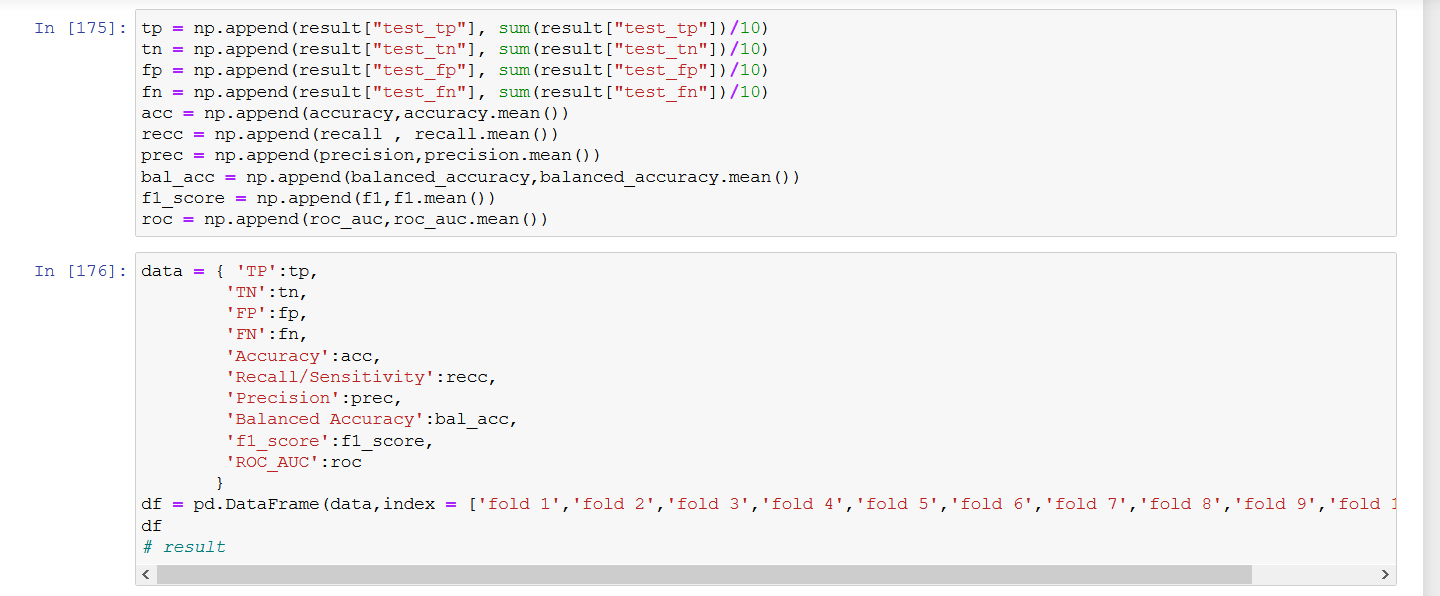


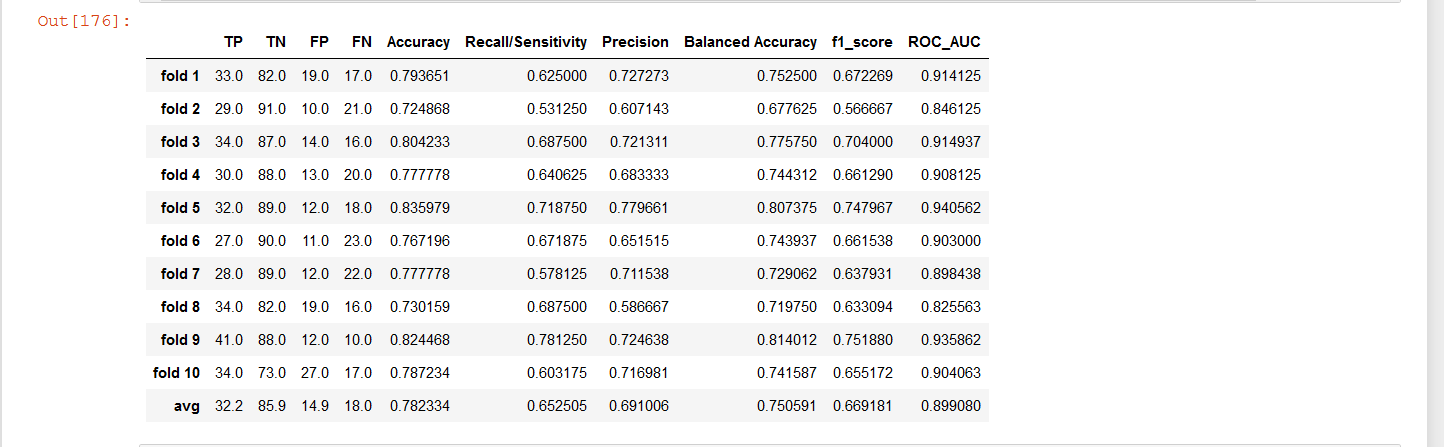












We have seen 3 different classifiers in this project:

* Random Forest
* Naïve Bayes
* KNN

We have also calculated various evaluating metrics that was asked for 10 folds in the cross validation.

According to the project Random Forest Classifier proves out to be the best among all of the three classifiers as its accuracy is 98 %.