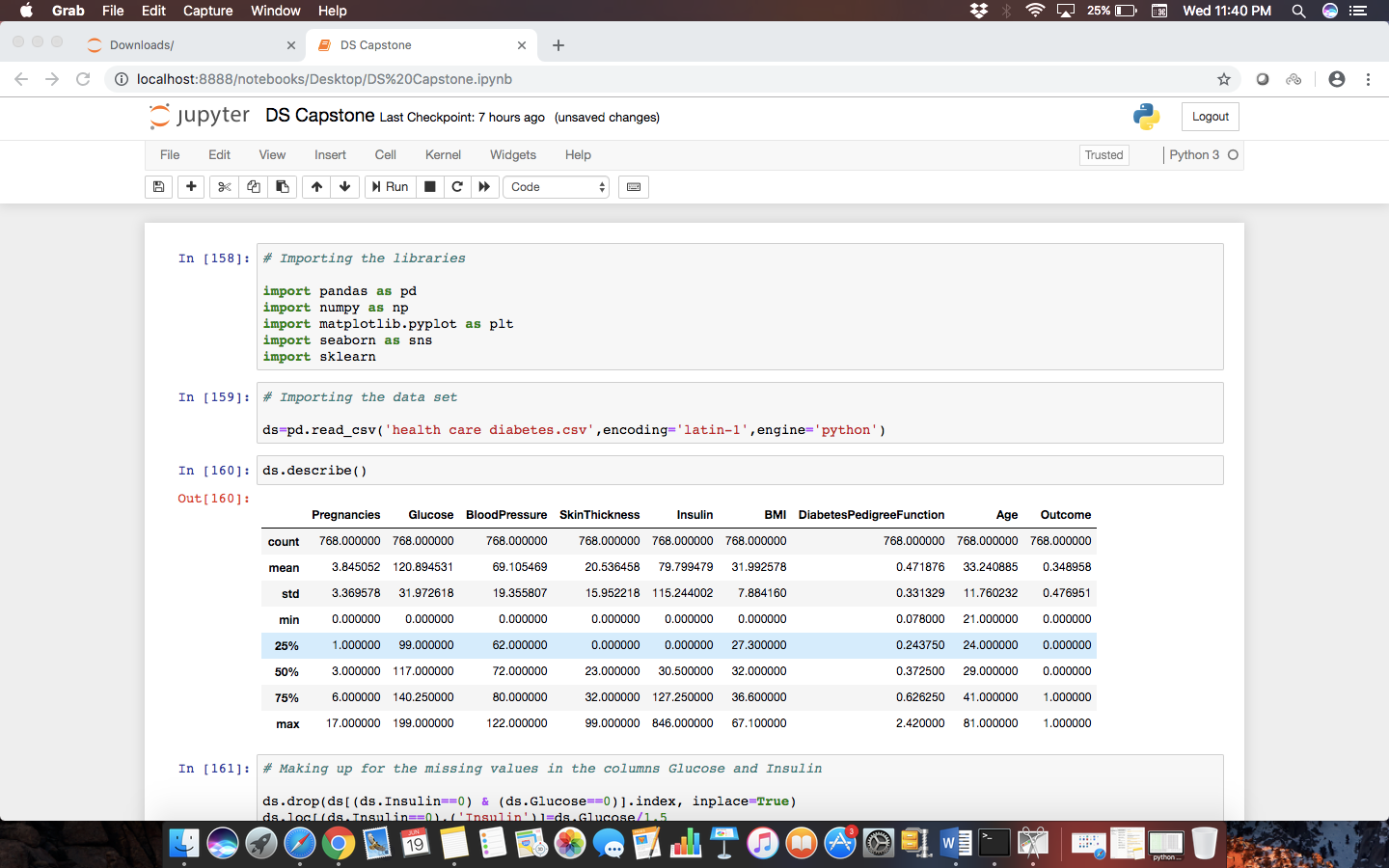
Data science capstone project 2 : Healthcare Industry

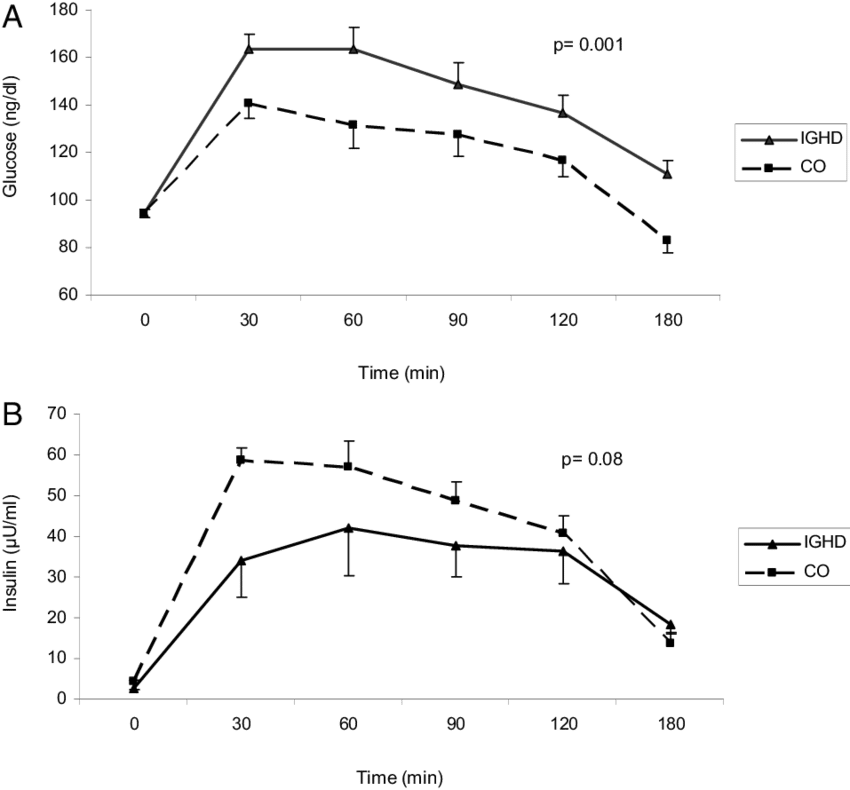


We have missing values in the columns of

Glucose, Insulin, Blood pressure, Skin Thickness and BMI

Step 1) Treat the missing values of the columns Glucose and Insulin.

We know that as the level of glucose increases, the level of insulin has to increase in order to control the increasing levels of blood glucose. Hence there is a linear relationship between these two entities.



So for the sake of simplicity, I decided to take the relationship as, Insulin = Glucose / 1.5

And correspondingly, Glucose = Insulin \* 1,5

The columns with both 0s were only few in number (4), hence they were deleted.

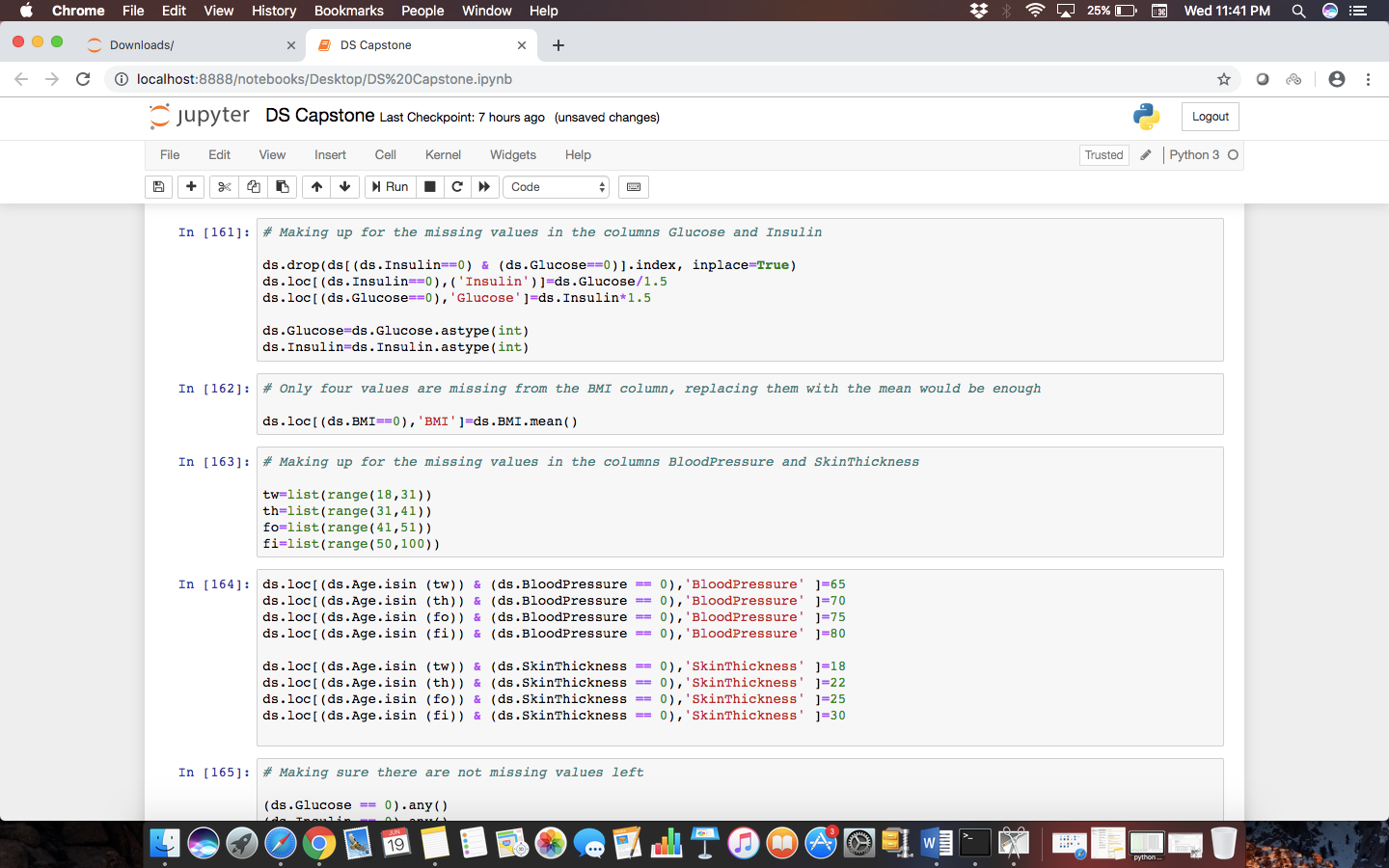
Step 2) The missing values of the column BMI.

Since there were only 5 missing values out of 760 values, I decided the to fill the missing values with the mean of the column.

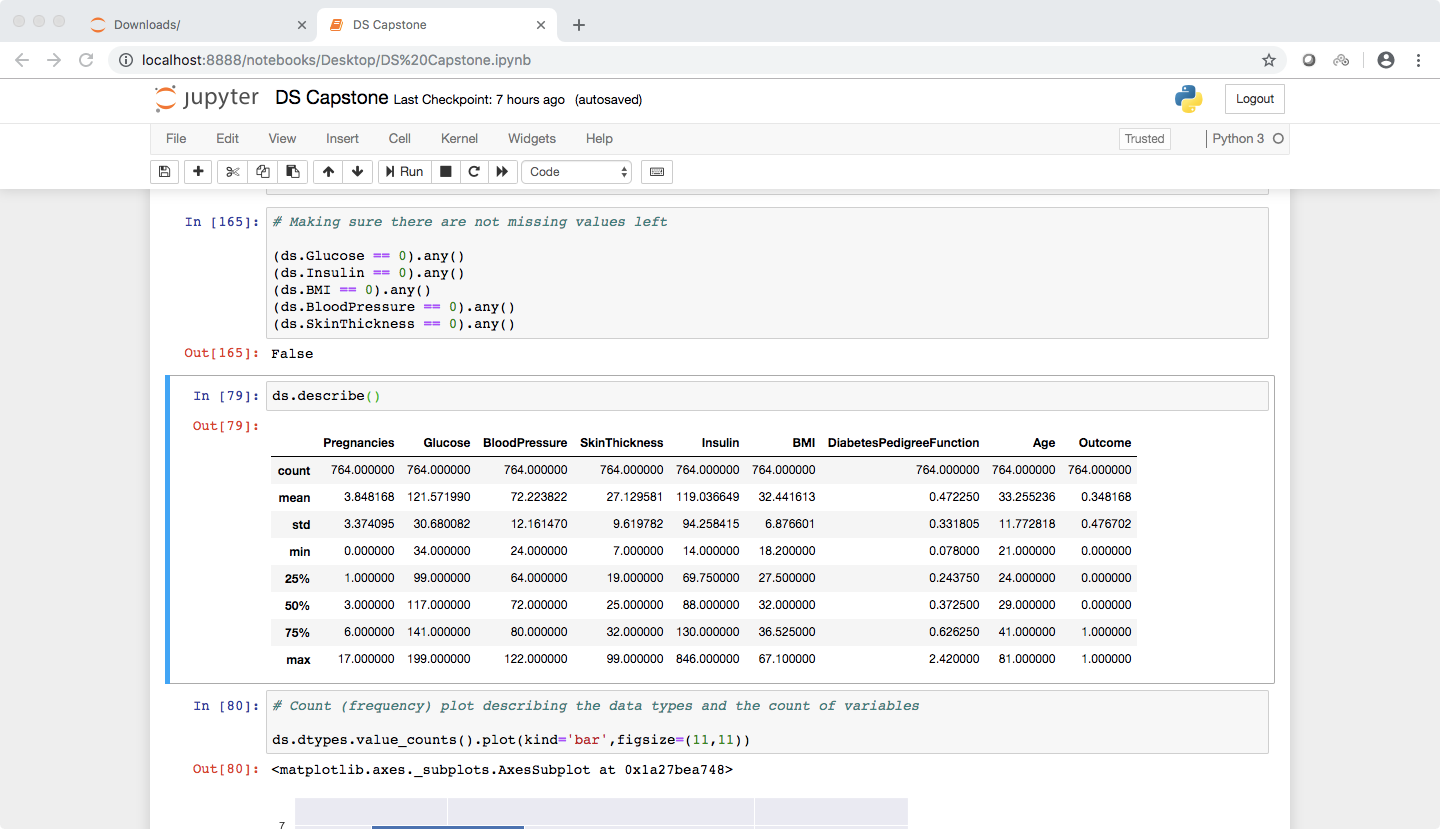
Step 3) The missing values of the column SkinThickness and BloodPressure.

Now as we know for a fact these columns depend largely on Age. Hence I decided to make use of the age column to fill the missing values of SkinThickness and BloodPressure.

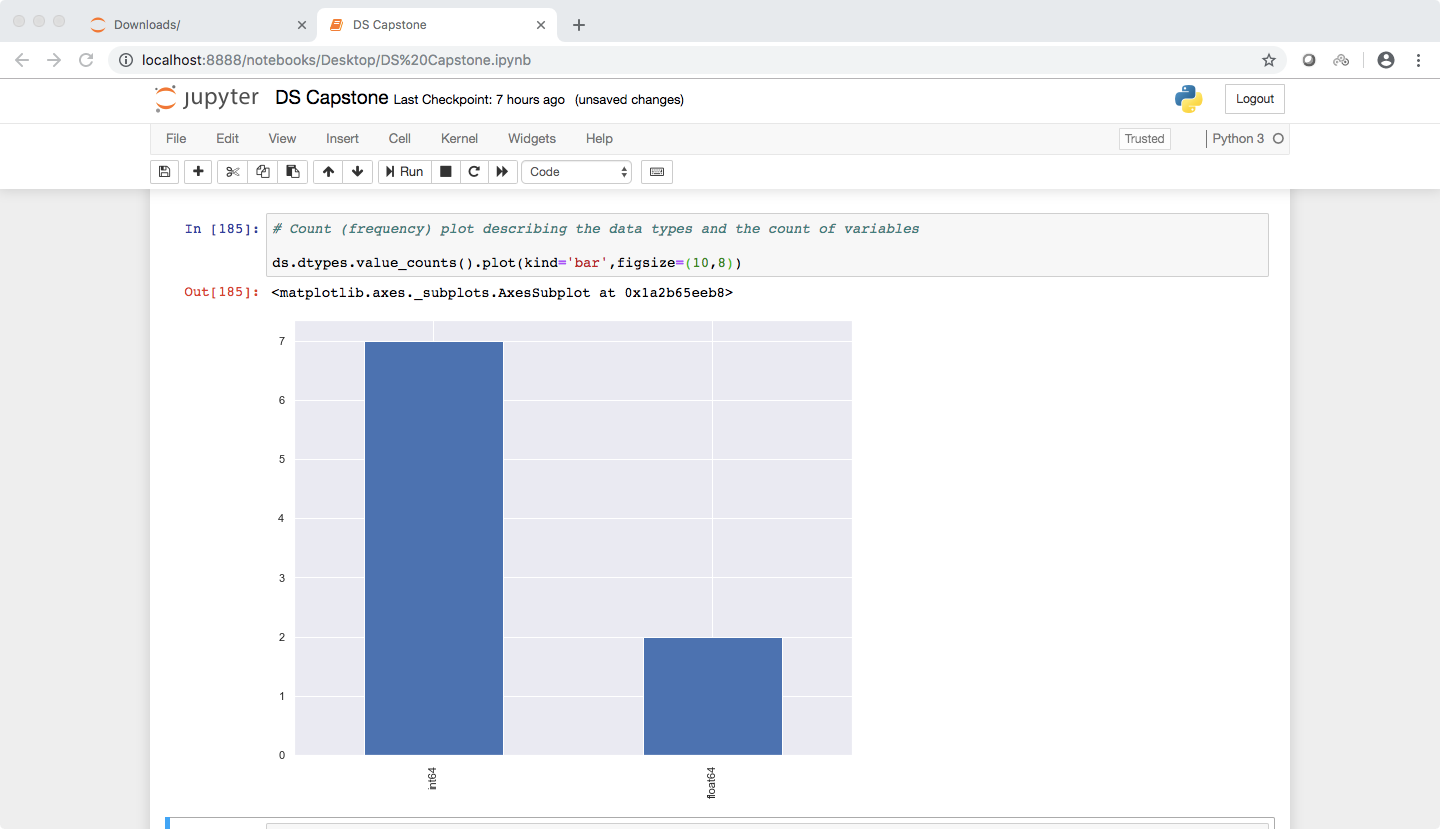
With the help of the Age column, I made 4 categories of ages and filled the BP and Skin columns correspondingly. For eg. A person of age 50-100 is likely to have high BloodPresssure(around 80) and skin thickness around 30 mm.



Thus the missing values were accounted for.

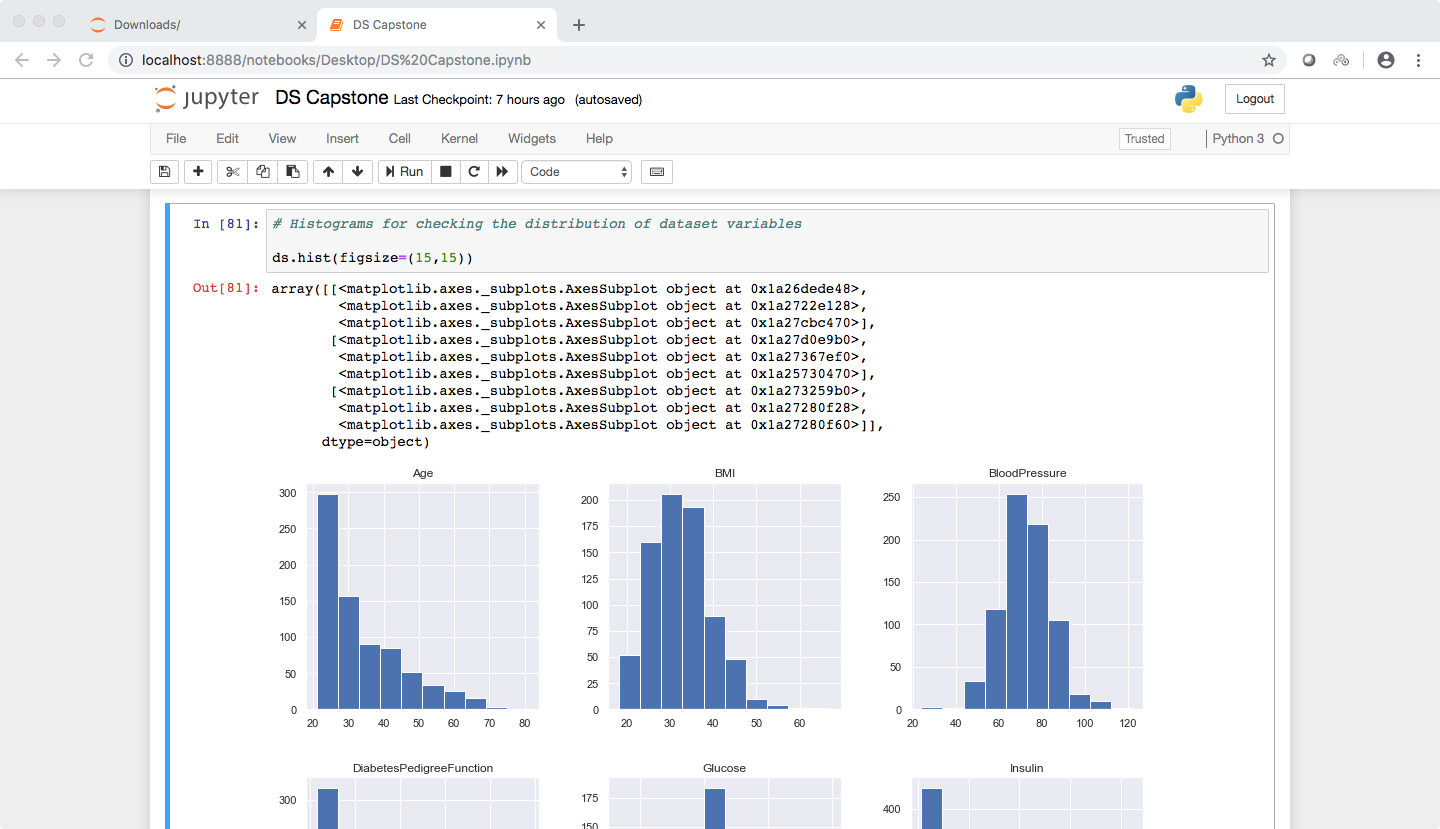


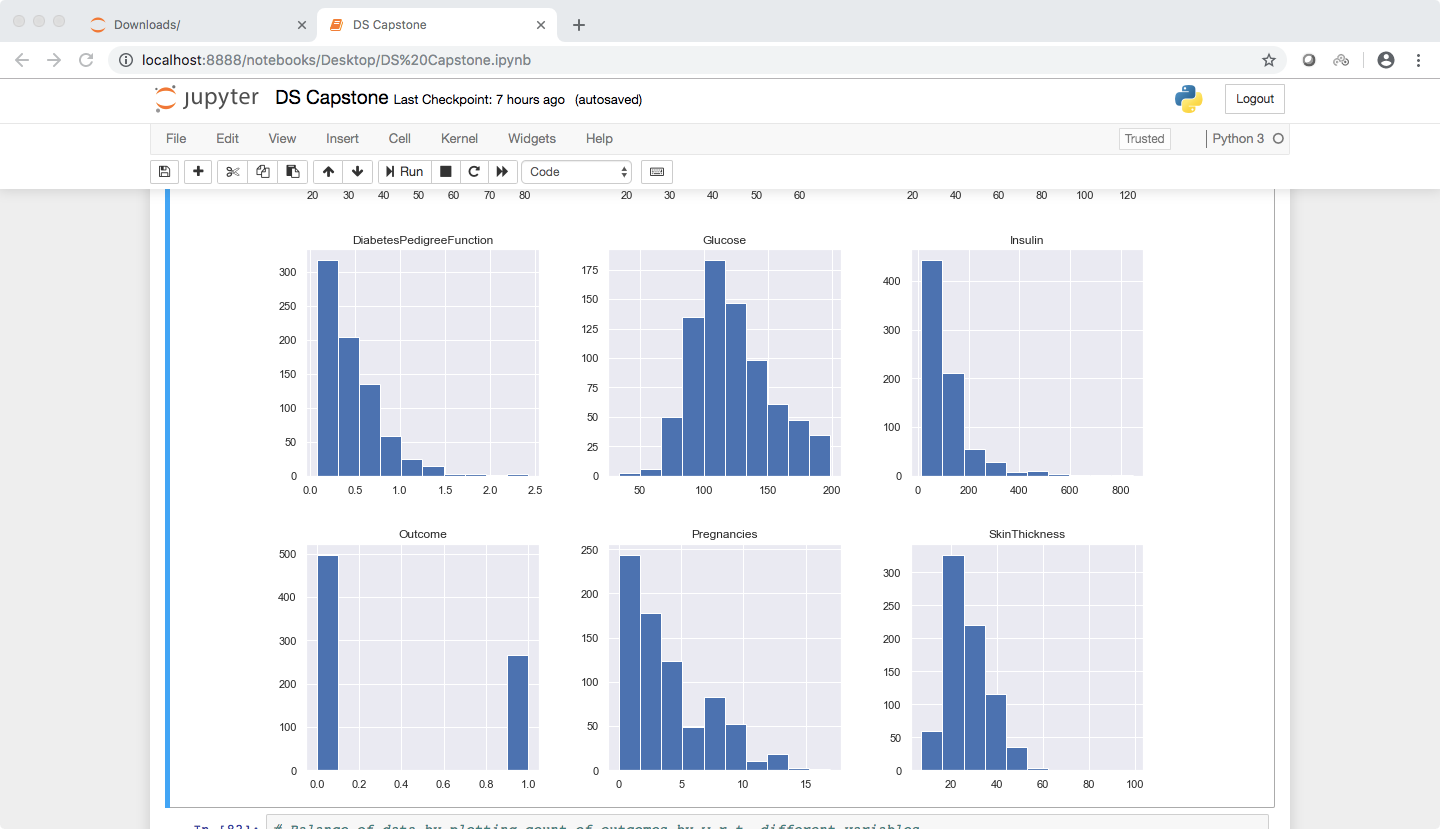
Plot describing the data types present in the data set and the count of variables corresponding to the data type.



Histograms of each Column in the dataset.

It can be seen that the majority of the population in the dataset falls in the 20-40 range of age. The majority of blood pressure level is in the 70s which is normal. And the Population with no diabetes is significantly larger than the one with diabetes. This is mainly because the population has young people as majority.





Balance of the data by plotting the count of outcomes by their value

1) For Age :-

a. Non diabetic women : As seen in the frequency distribution, the women with no diabetes lie mostly in the 20-35 age range.

b. Diabetic women : As seen the diabetic population is present in evenly and mostly in the 30-50 age range.

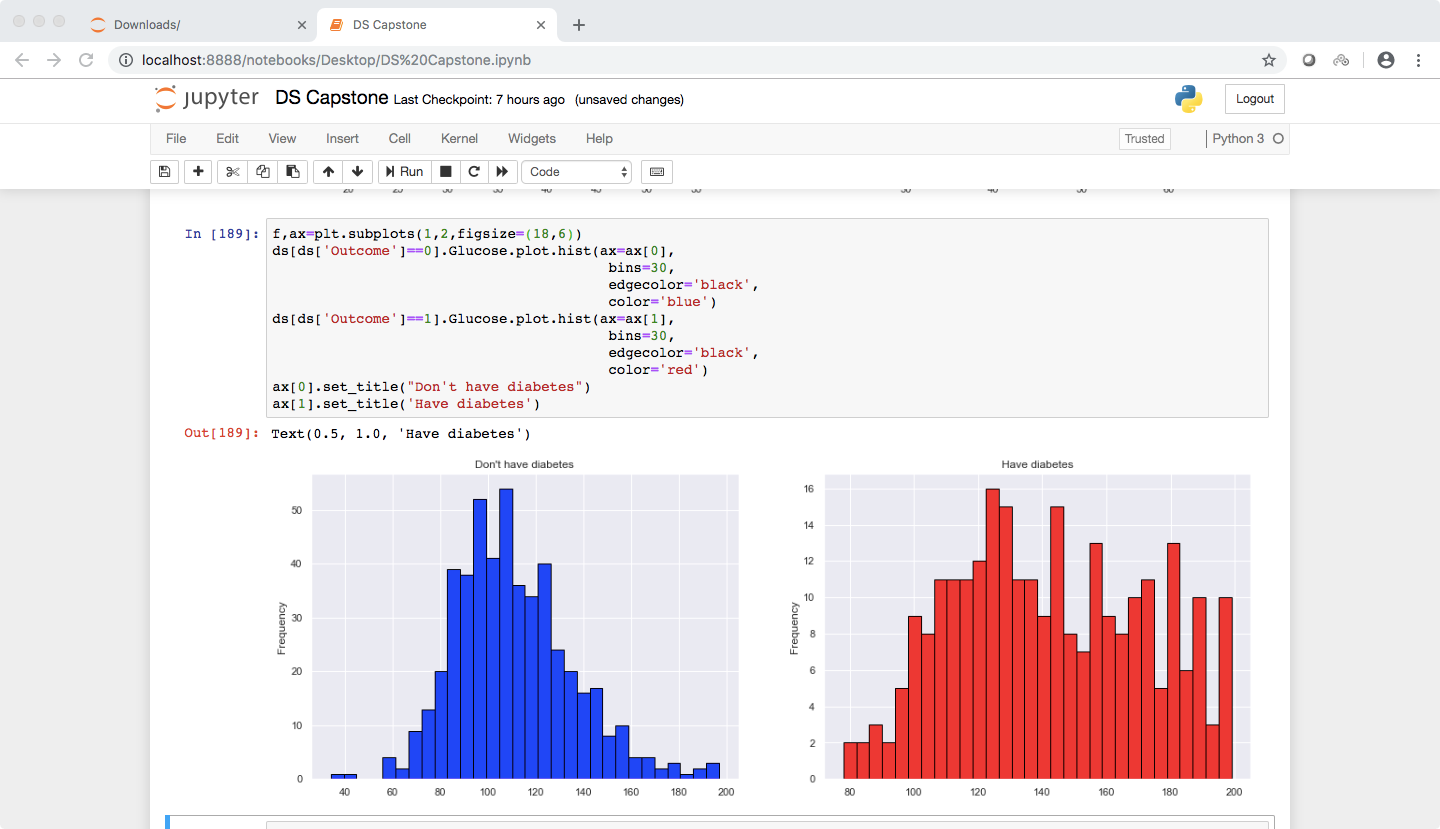


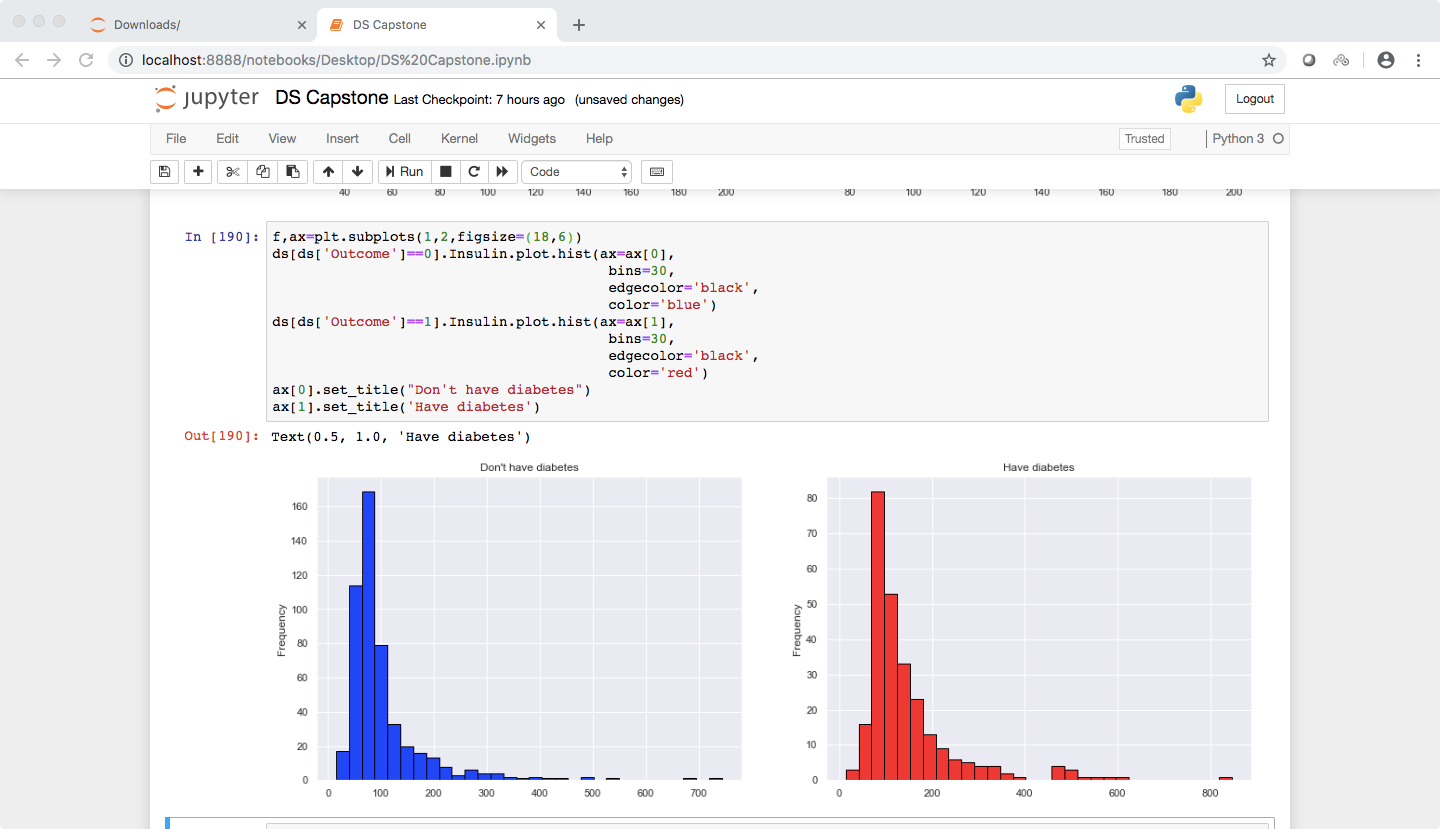
For BMI :- Nothing deducible here.



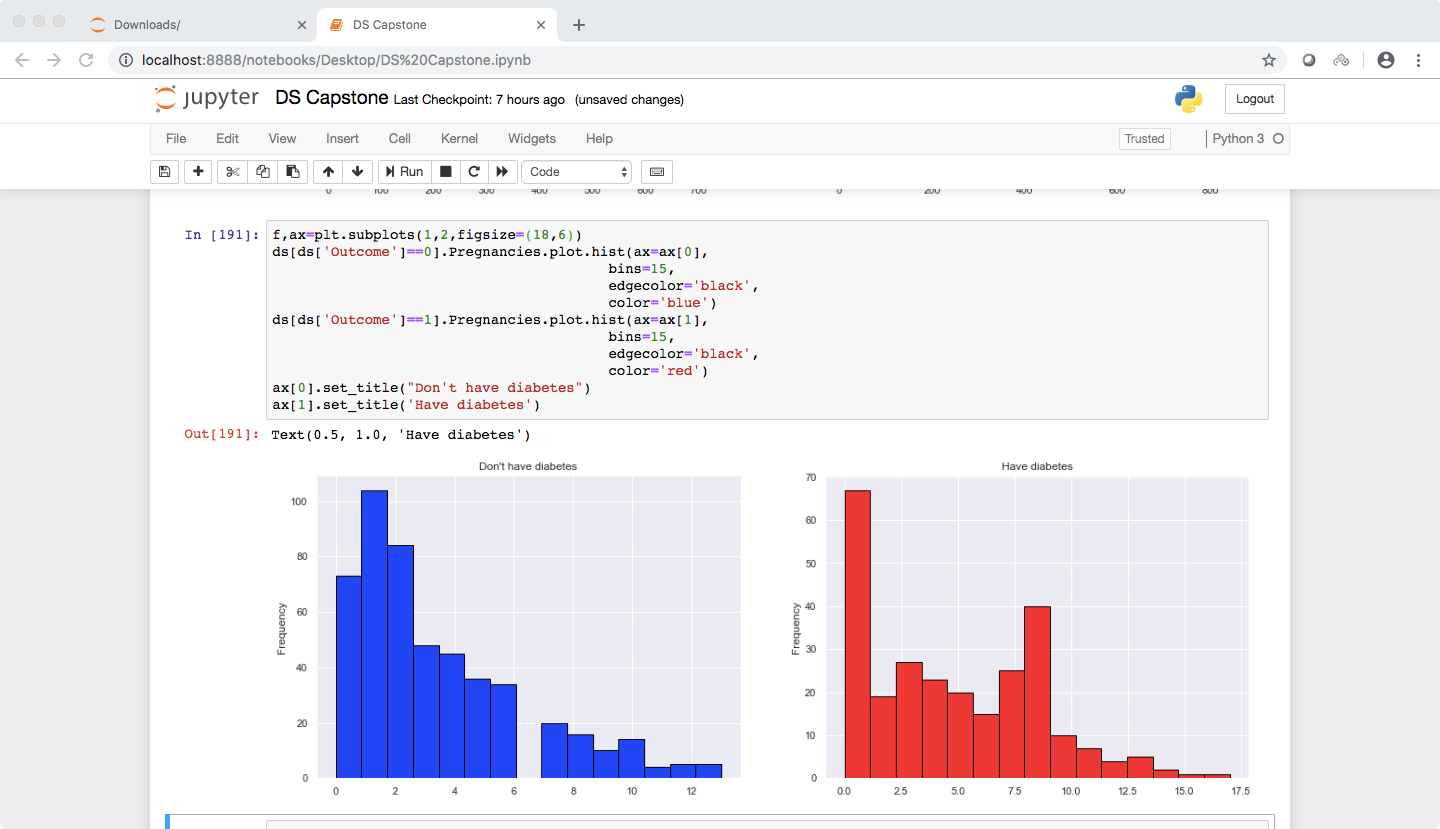
For Glucose:-

It can be seen in the diabetic graph that higher the level of glucose(120-200) higher are the count of diabetic patients.





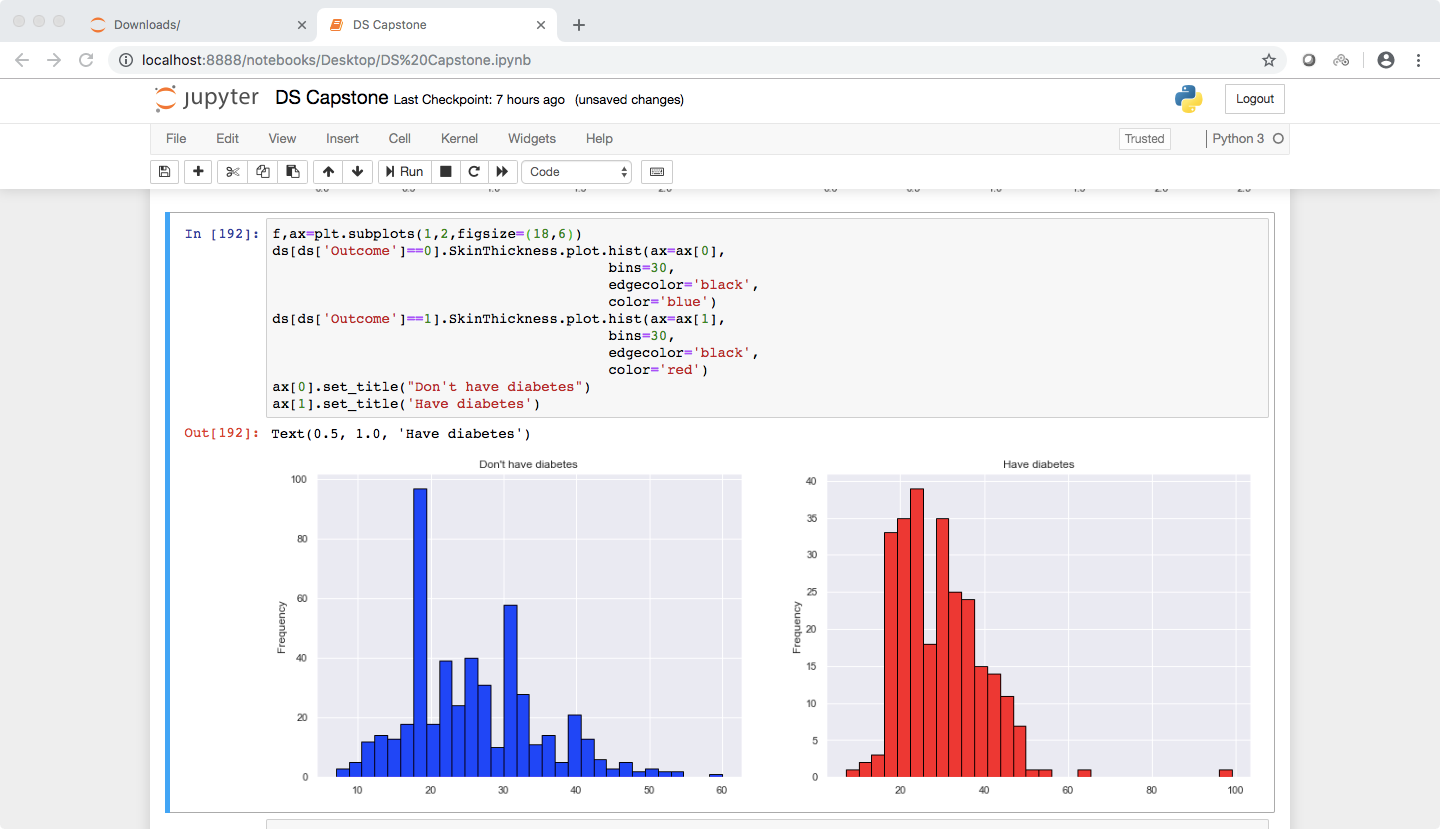
For Pregnancies :- It can be seen that more the number of pregnancies, more are the chances of getting diabetes.



For Skin Thickness :-

Diabetic women : In the group of non diabetic women, the women with less skin thickness are observed to have diabetes.

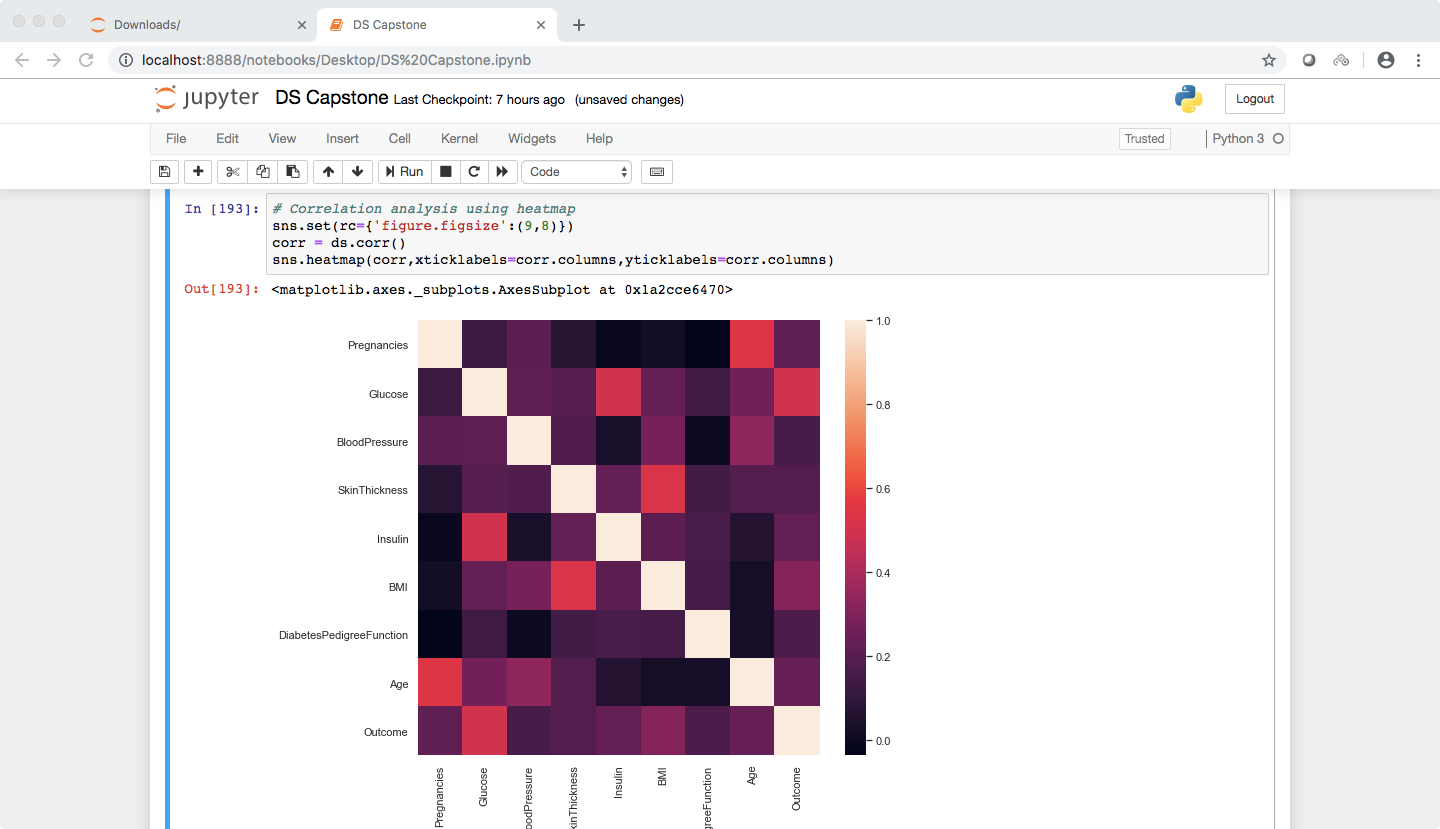
While the graph for non diabetic women doesn’t seem to be that informative.



The correlation analysis with heat map.

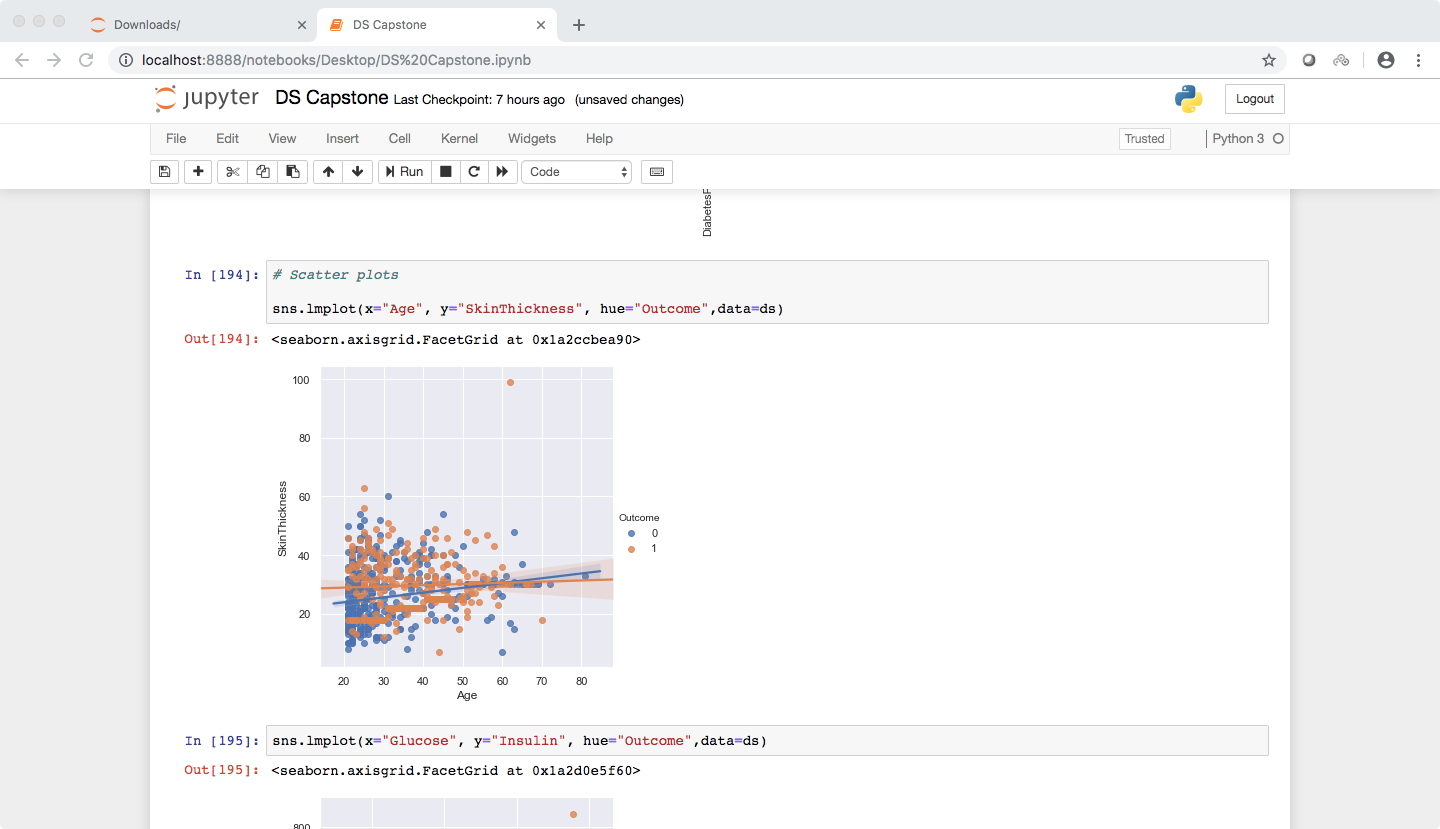
It can be seen that the columns Glucose,BMI,Age,Insulin,Pregnancies have varying influences on the outcome column.

While the columns like Pedgreefunction BloodPressure don’t really have any affect on it.

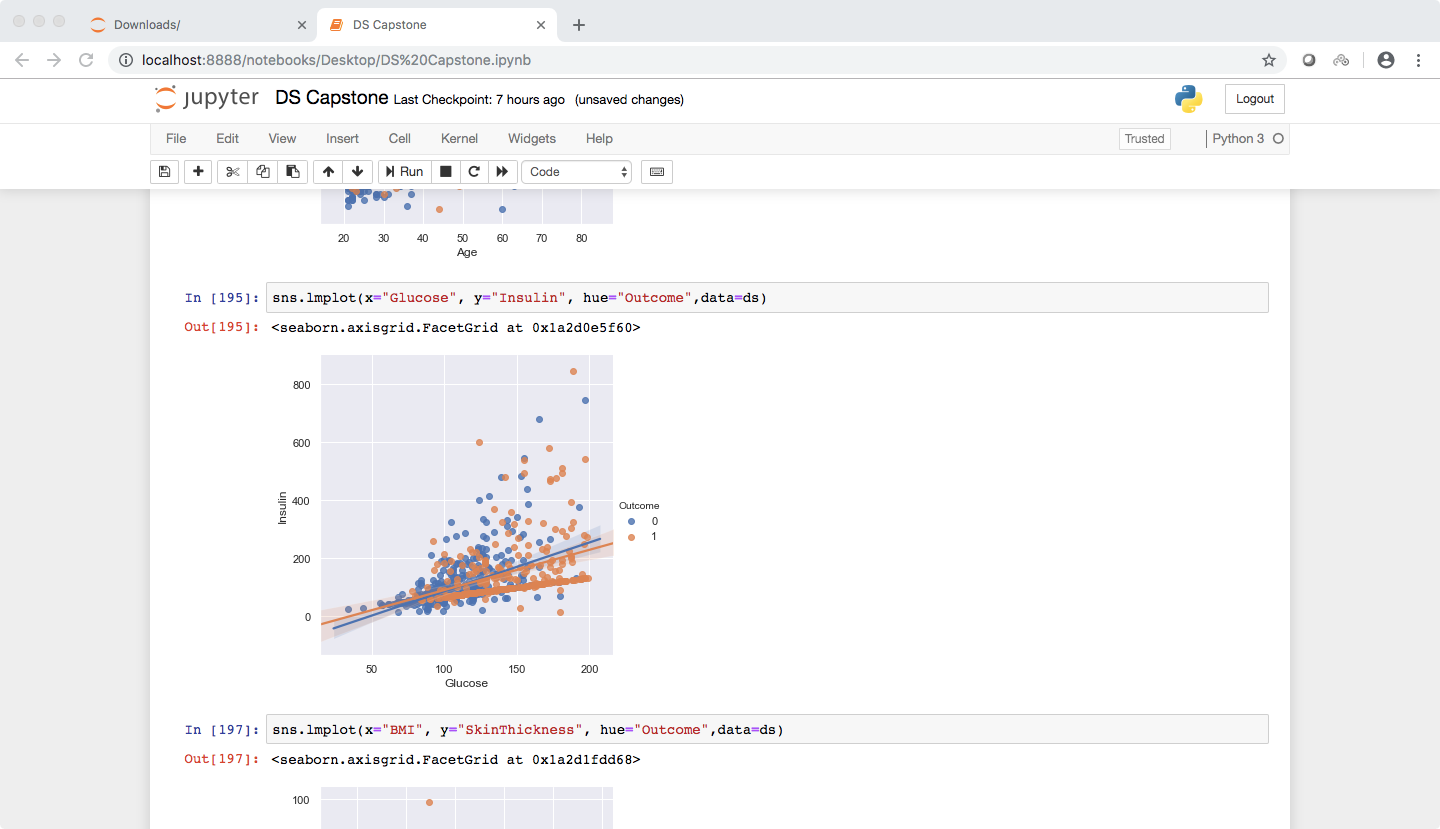


Scatter plots

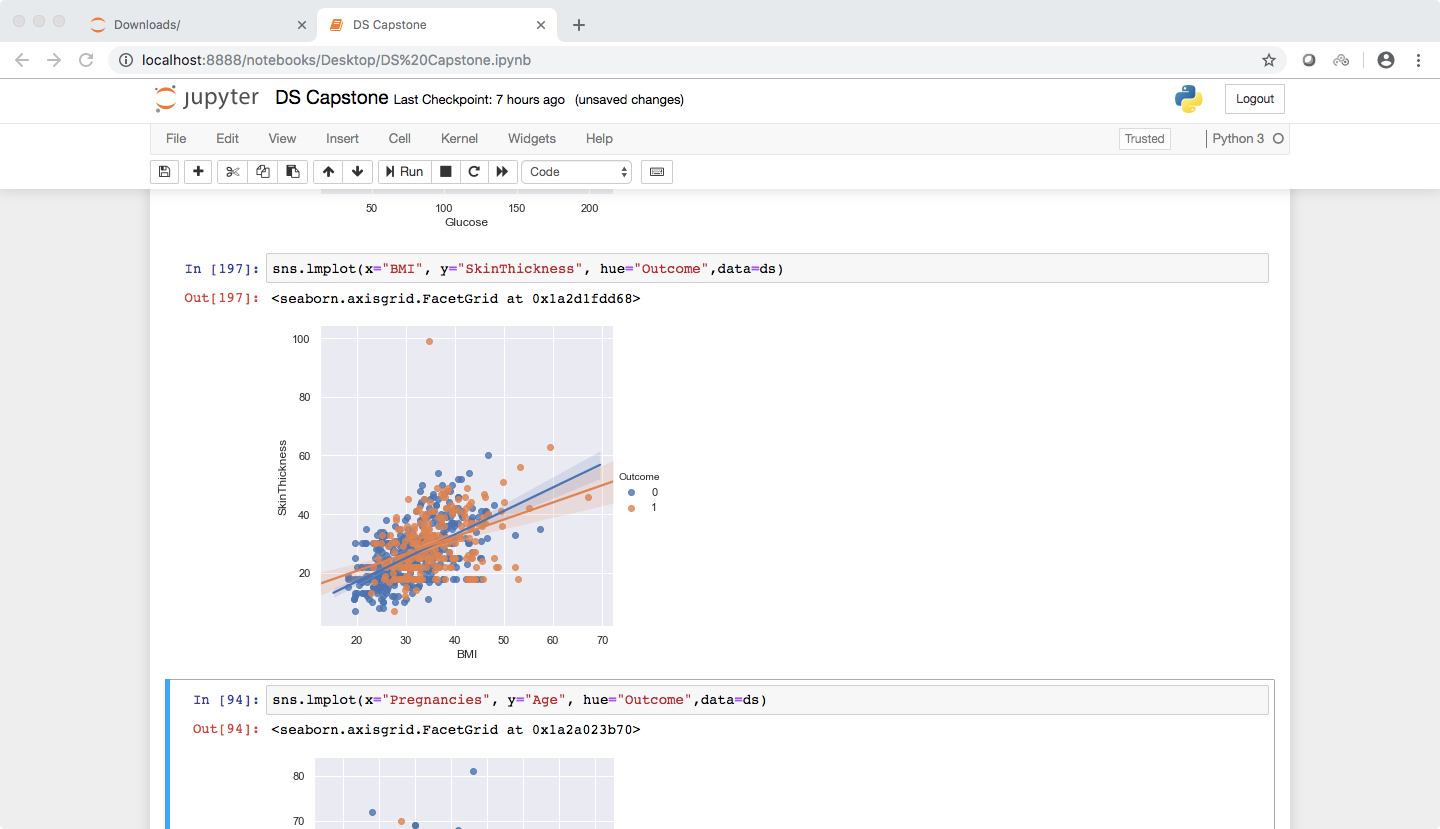
It can be seen that there is correlation between the age and skin thickness. There are no significant amount of outliers except 4-5.

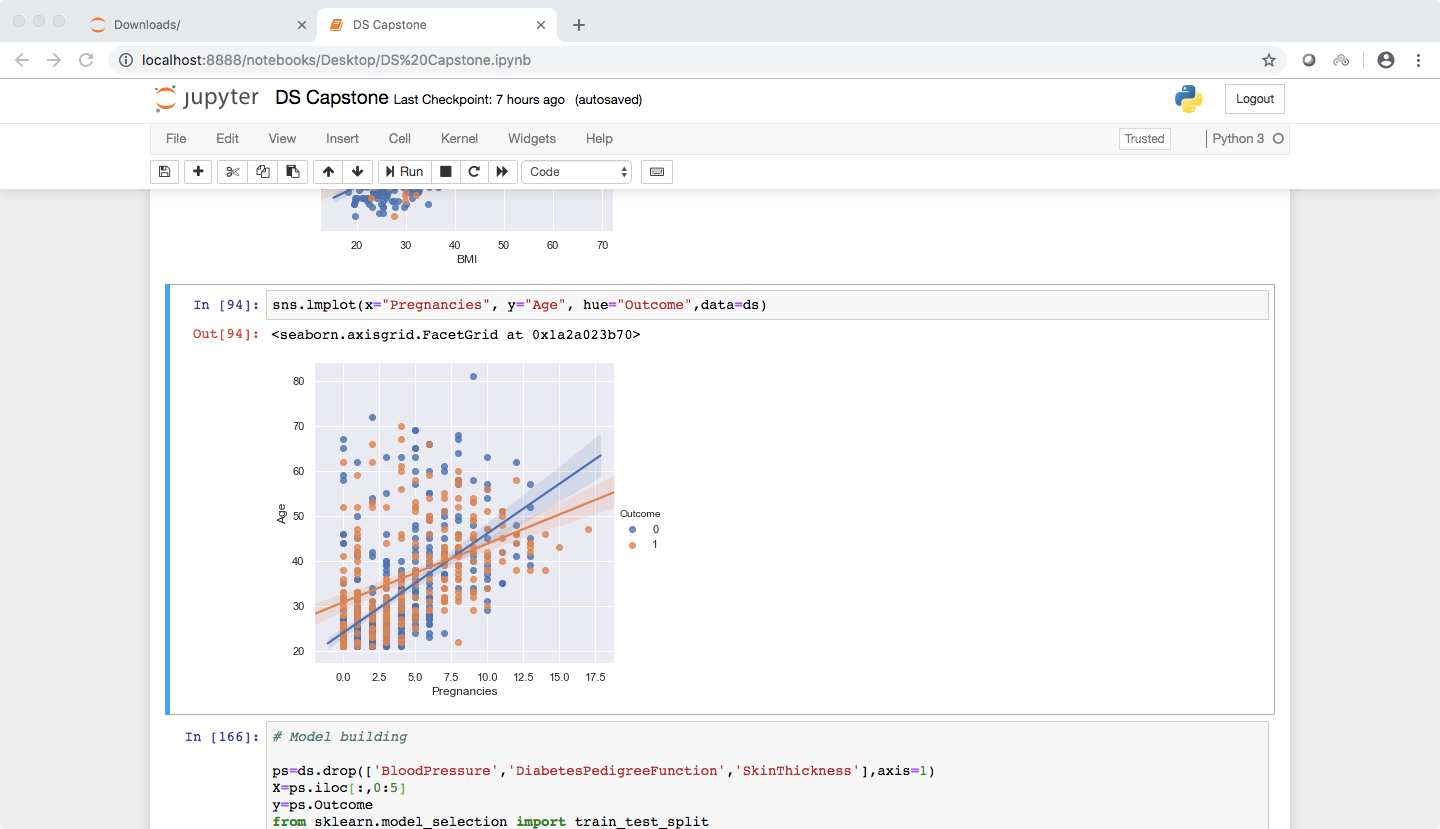


It can be seen that there is correlation between the Glucose and Insulin. However there are outliers present too but not enough to make any significant difference.



BMI and skin thickness have very strong correlation.



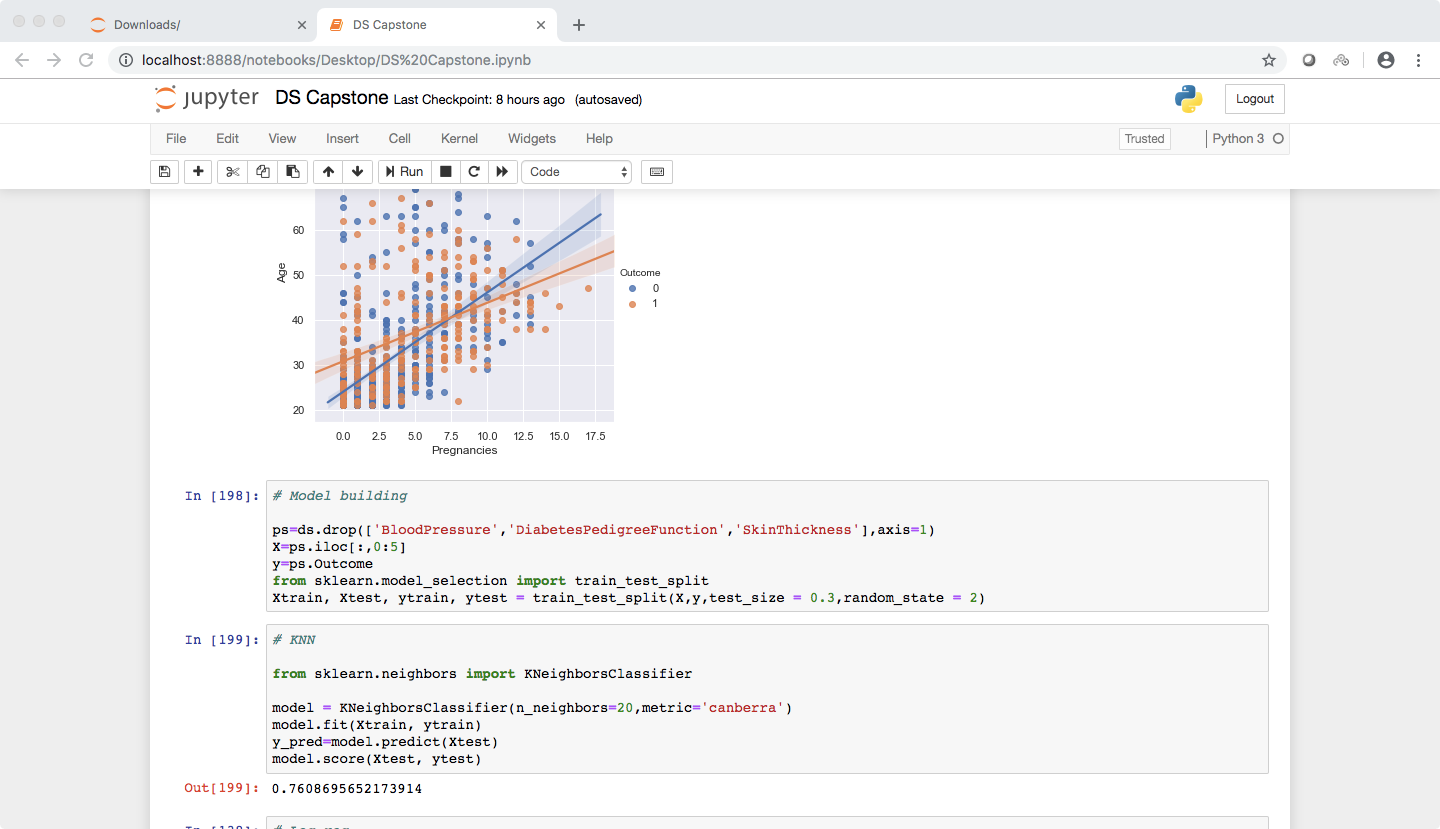


Model building :-

The columns of Blood Pressure, Pedigree Function and Skin Thickness don’t really have to be taken into consideration while predicting whether a person has diabetes or not (as seen in the heatmap) so I decided to drop them.

The first model used is KNN. Through some trial and error I found that the model gave the best result for 20 neighbors and metric = canberra.

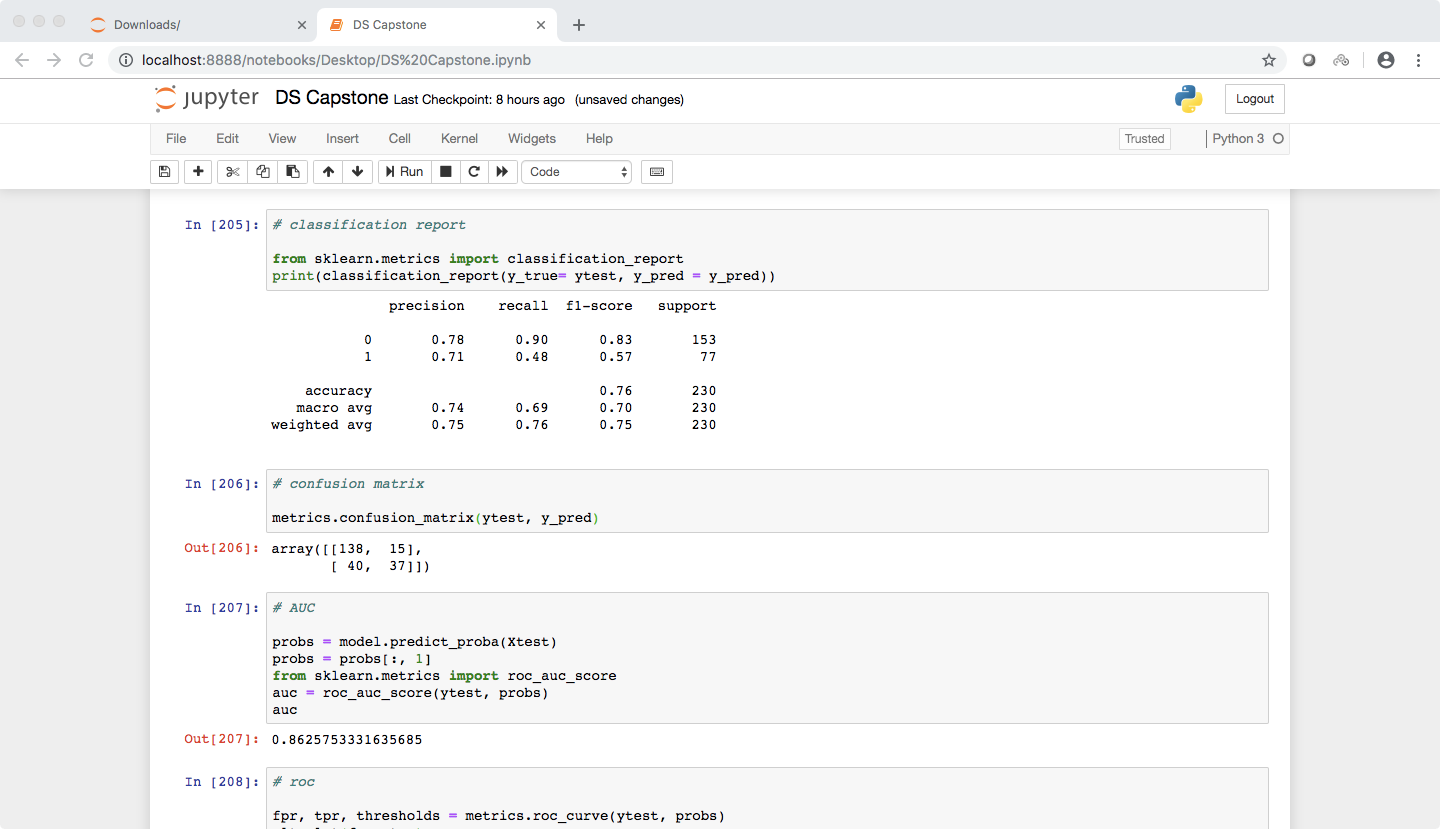
The accuracy score was 76.08 %.



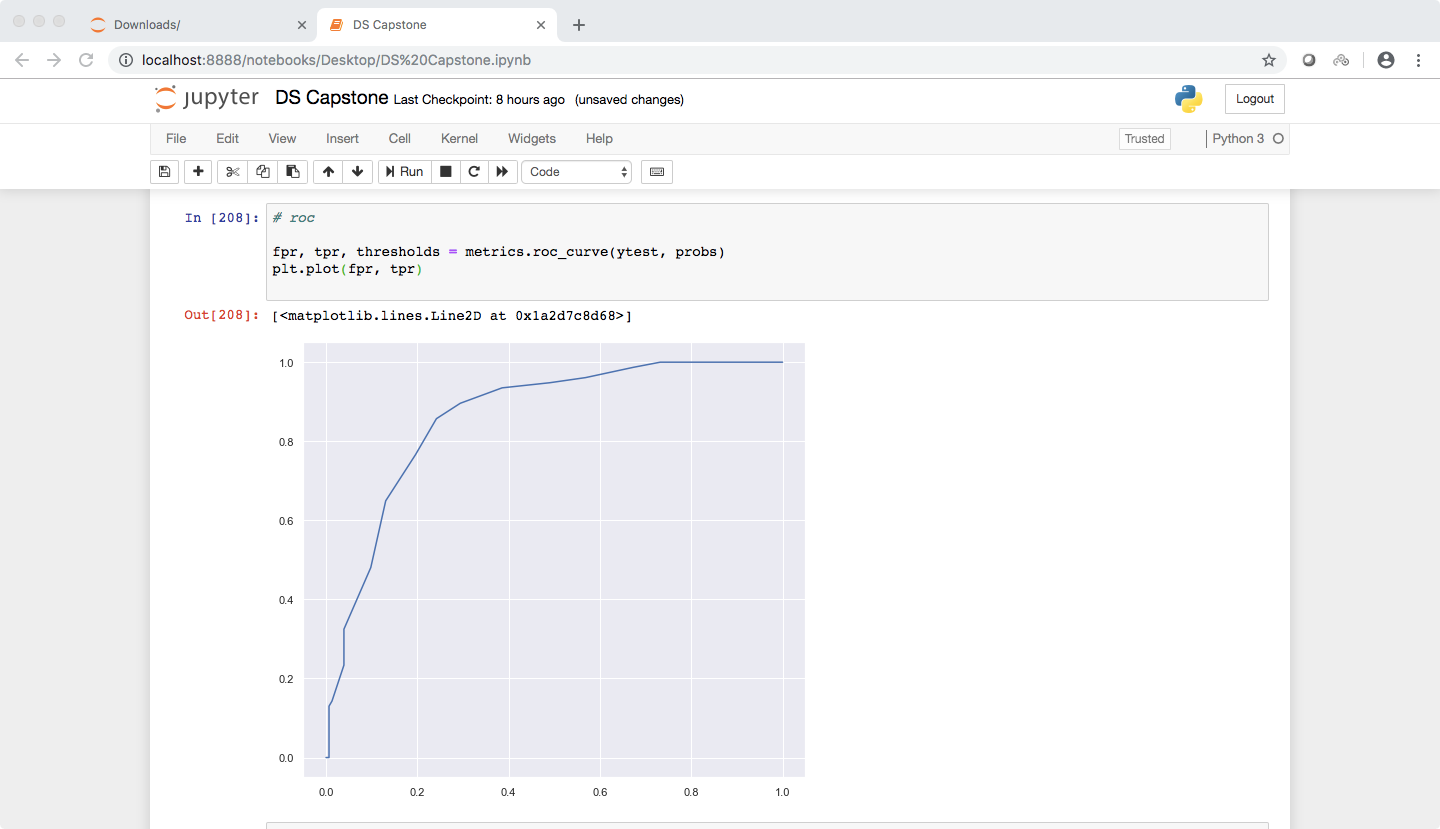
As seen in the classification report below, It had a specificity of 90 % and sensitivity as 48 %

And according to the confusion matrix, it had TP=138 TN=37 FP=15 FN=40

Had an AUC score of 86.25 %

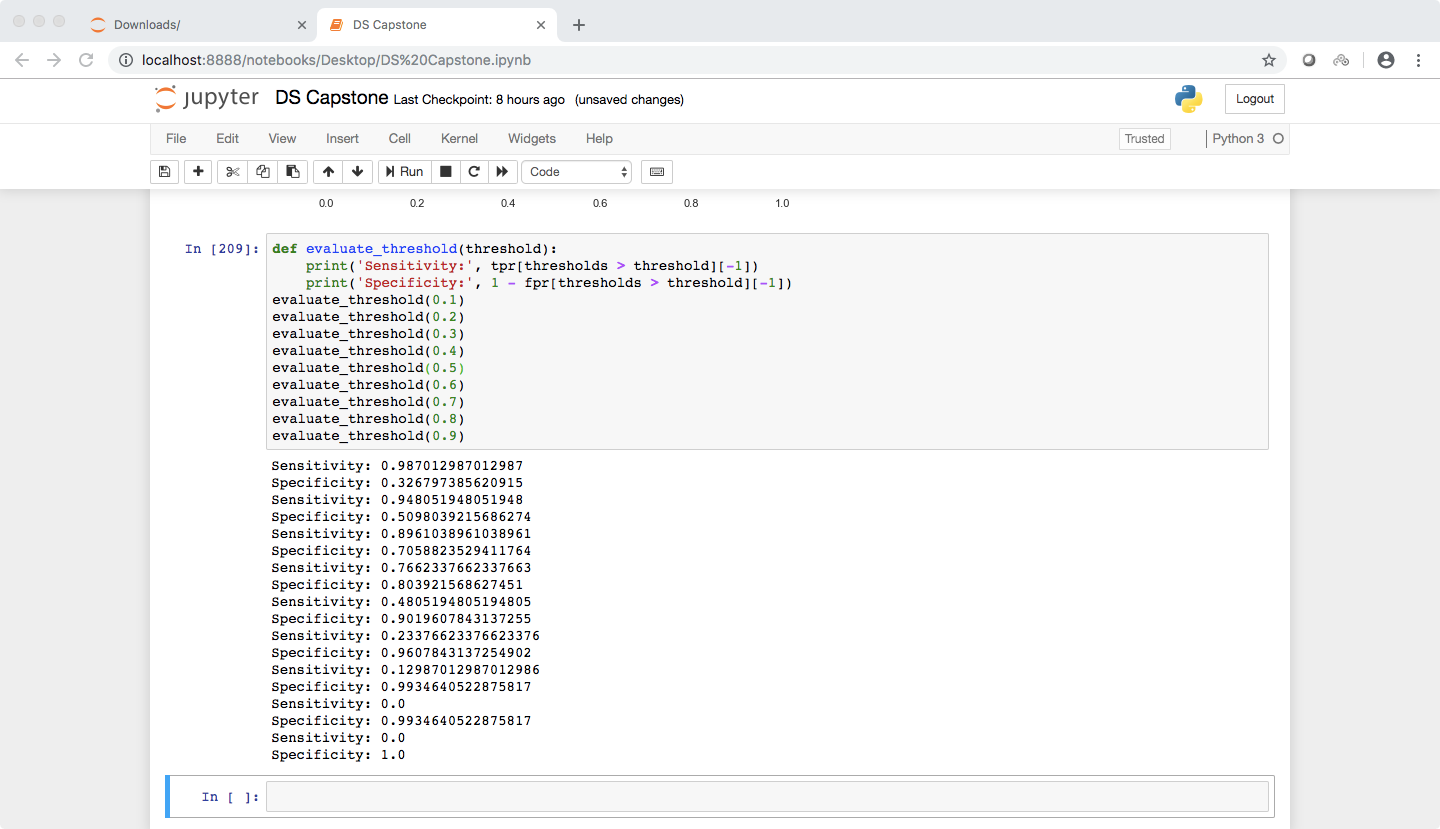


Below is the roc curve

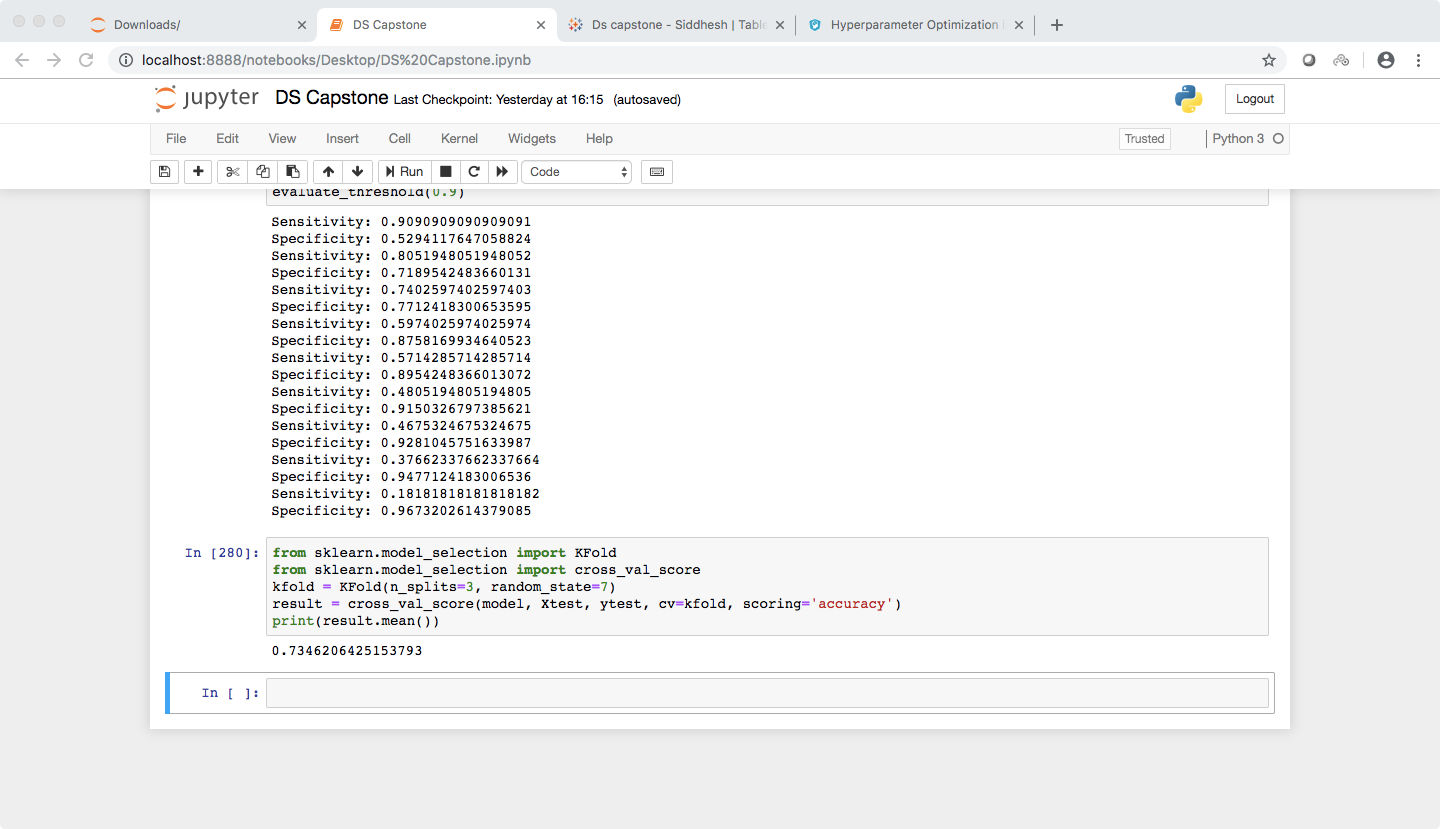


At a threshold of 0.3 it gave specificity = 89 % and sensitivity = 70 %.

So 30 % was the optimal threshold since it is a medical case and it would be better if it had low False Negatives.

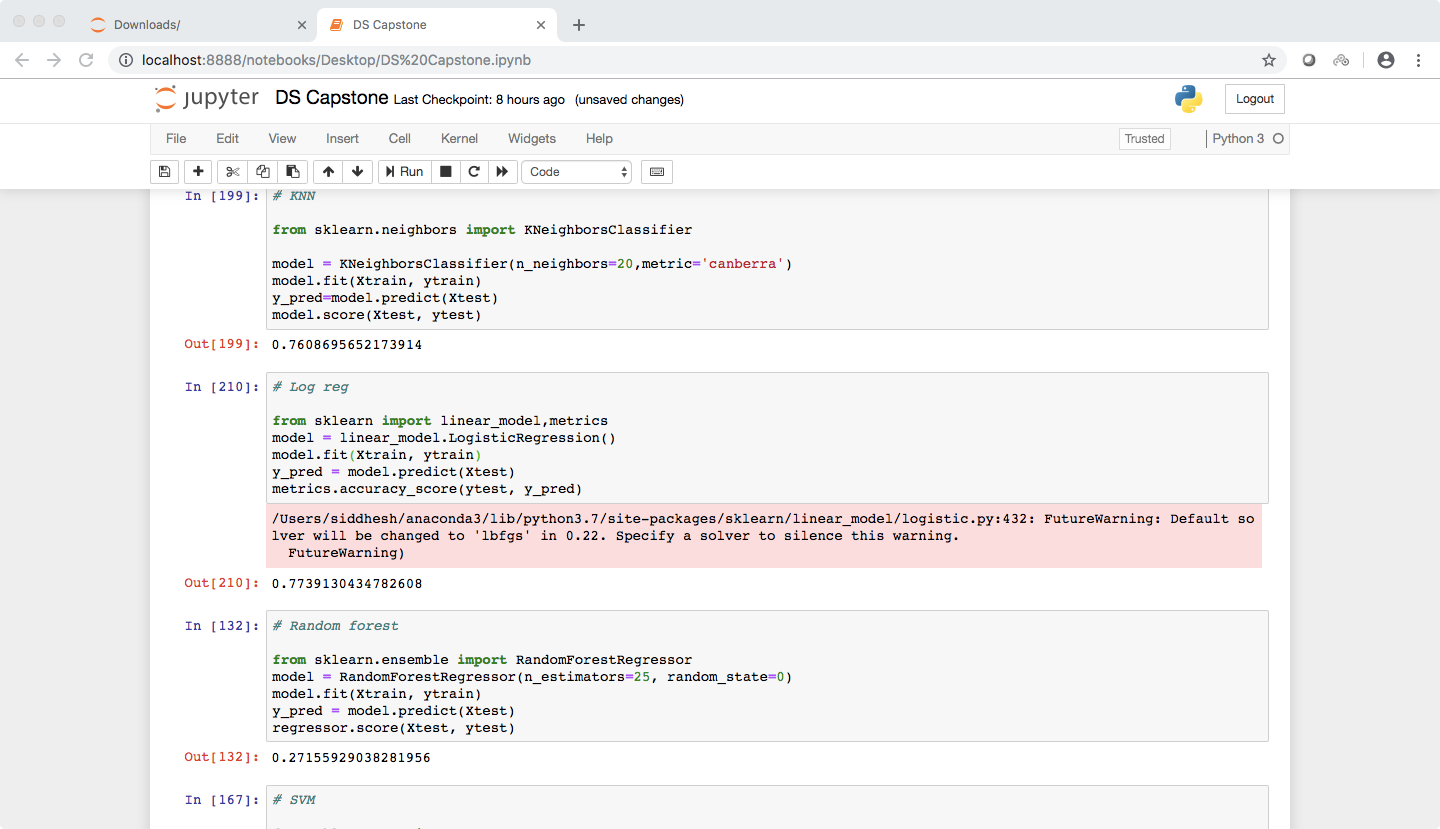


It had a cross validation score of 73.46 %



The second model I used as Logistic regression.

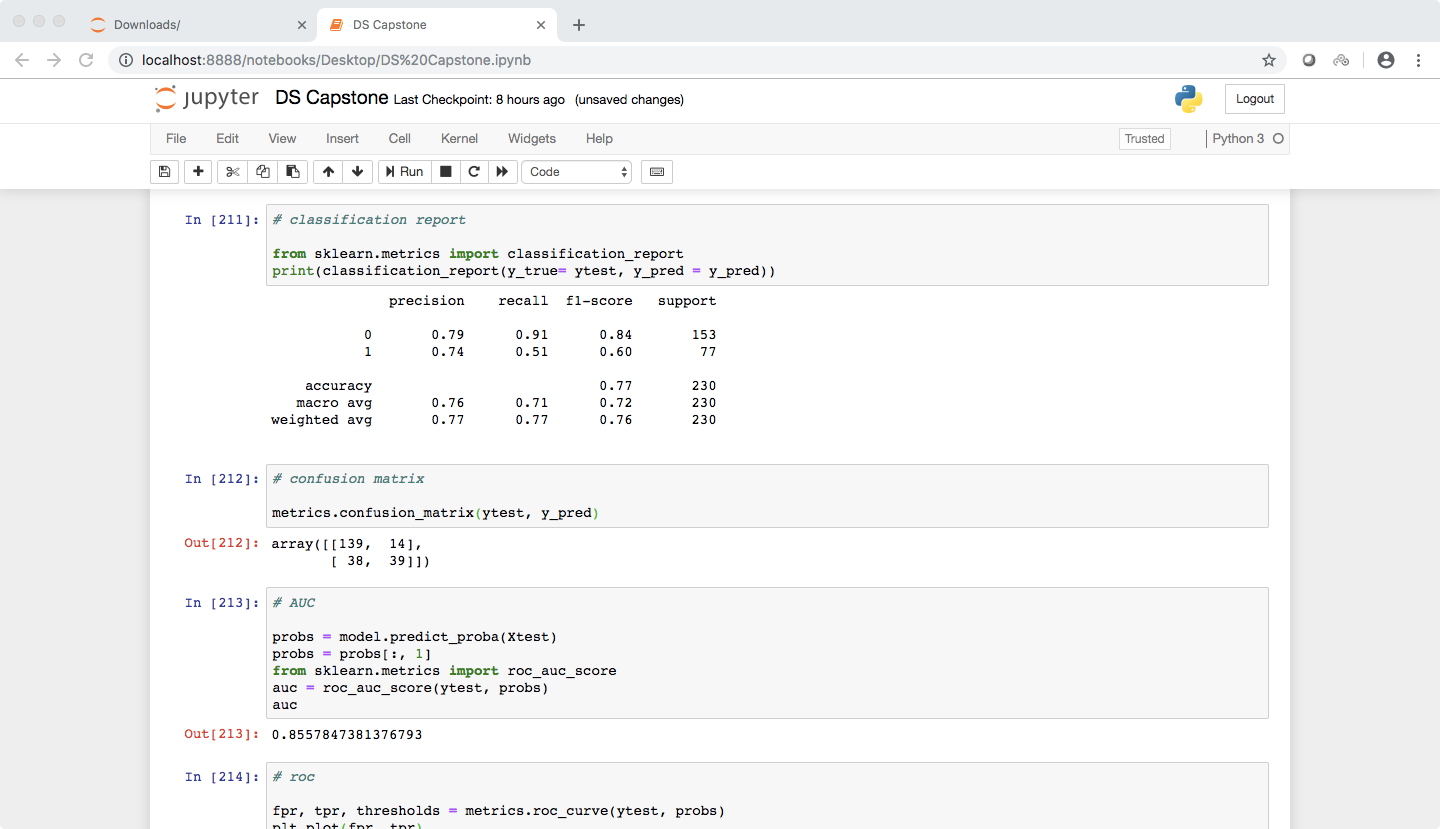
The accuracy score was 77.39 %.



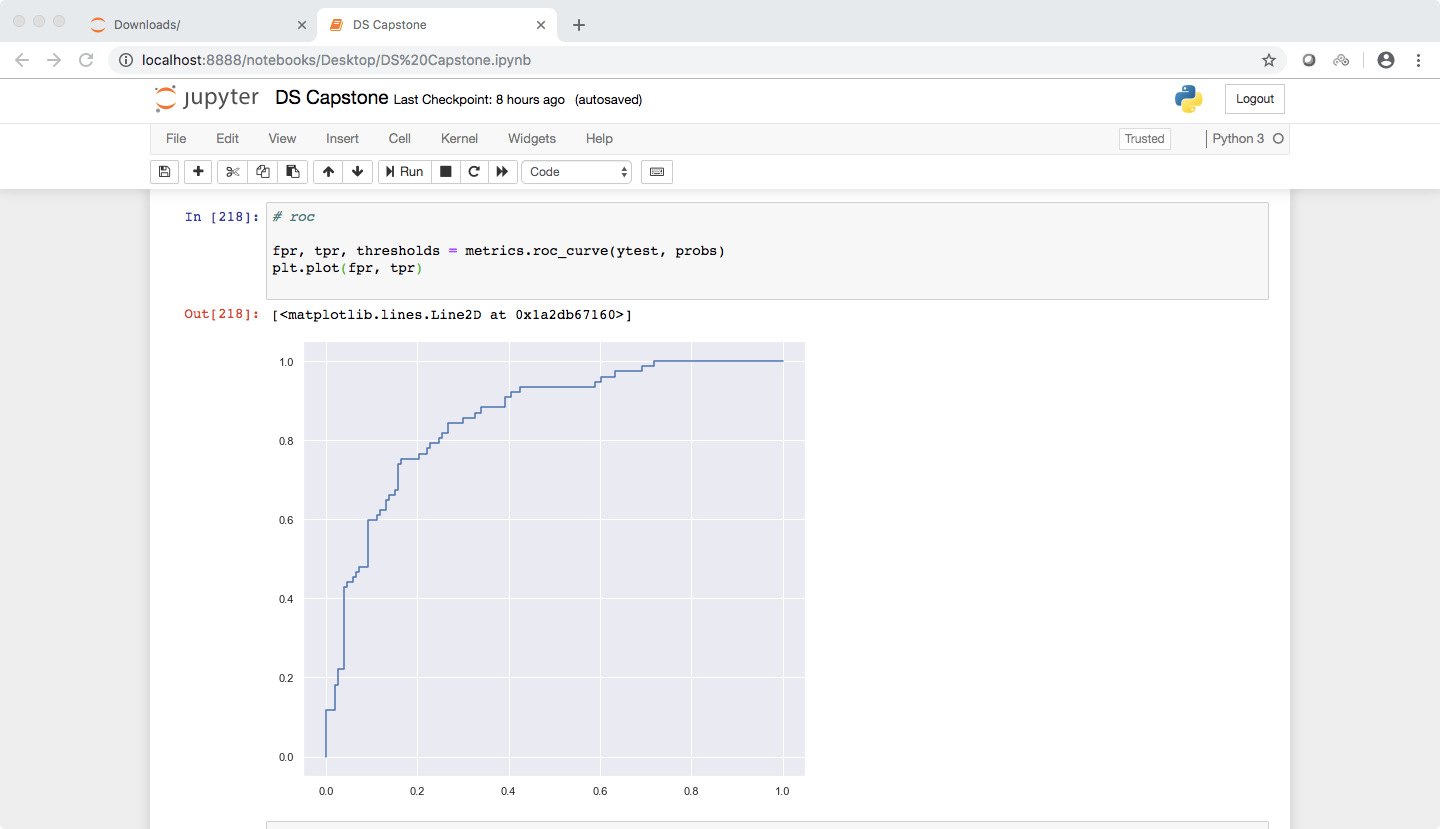
As seen in the classification report below, It had a specificity of 91 % and sensitivity as 51 %

And according to the confusion matrix, it had TP=139 TN=39 FP=14 FN=38

Had an AUC score of 85.55 %

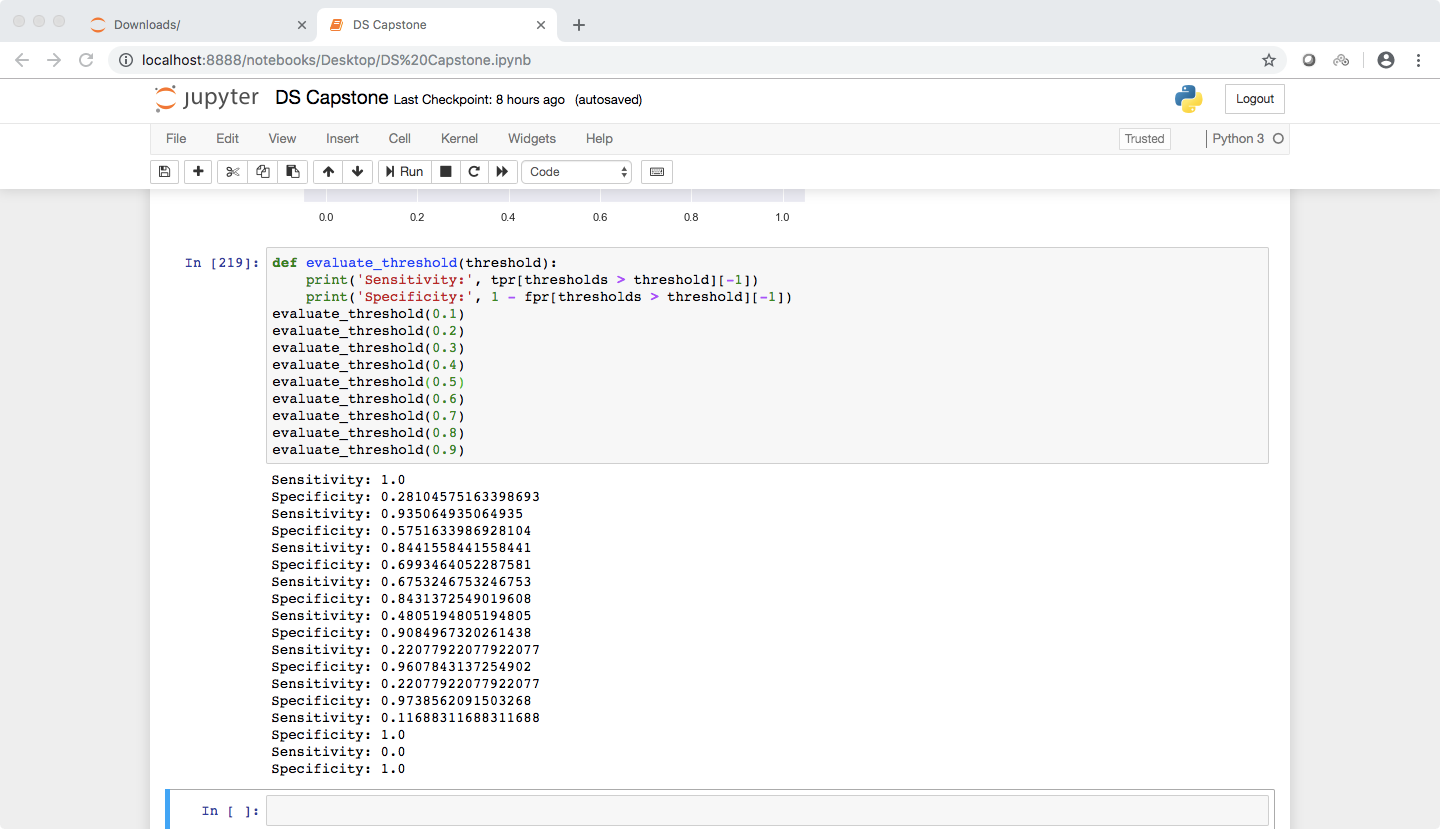


Below is the roc curve

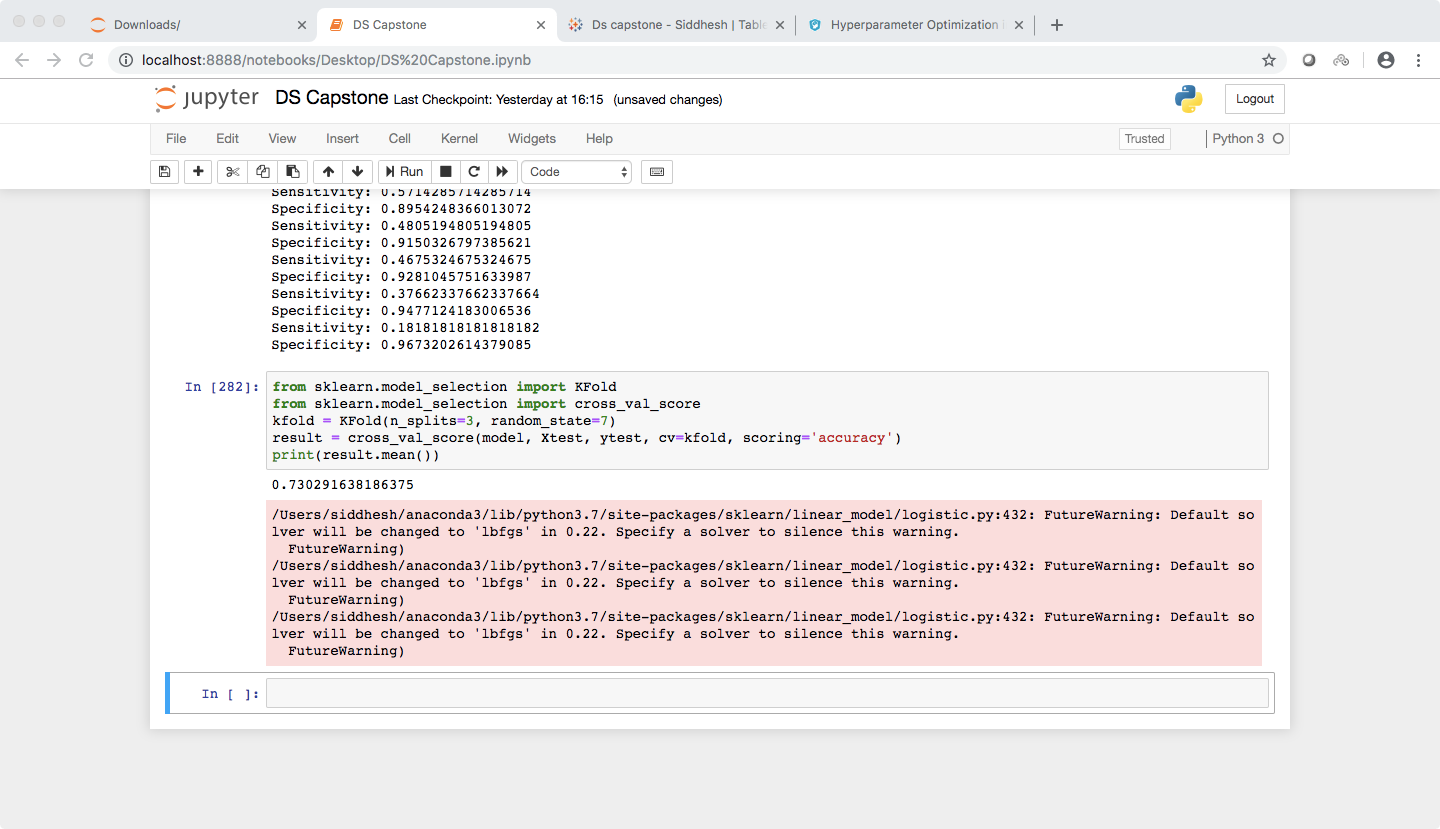


At a threshold of 0.3 it gave specificity = 84 % and sensitivity = 69 %.

So 30 % was the optimal threshold.

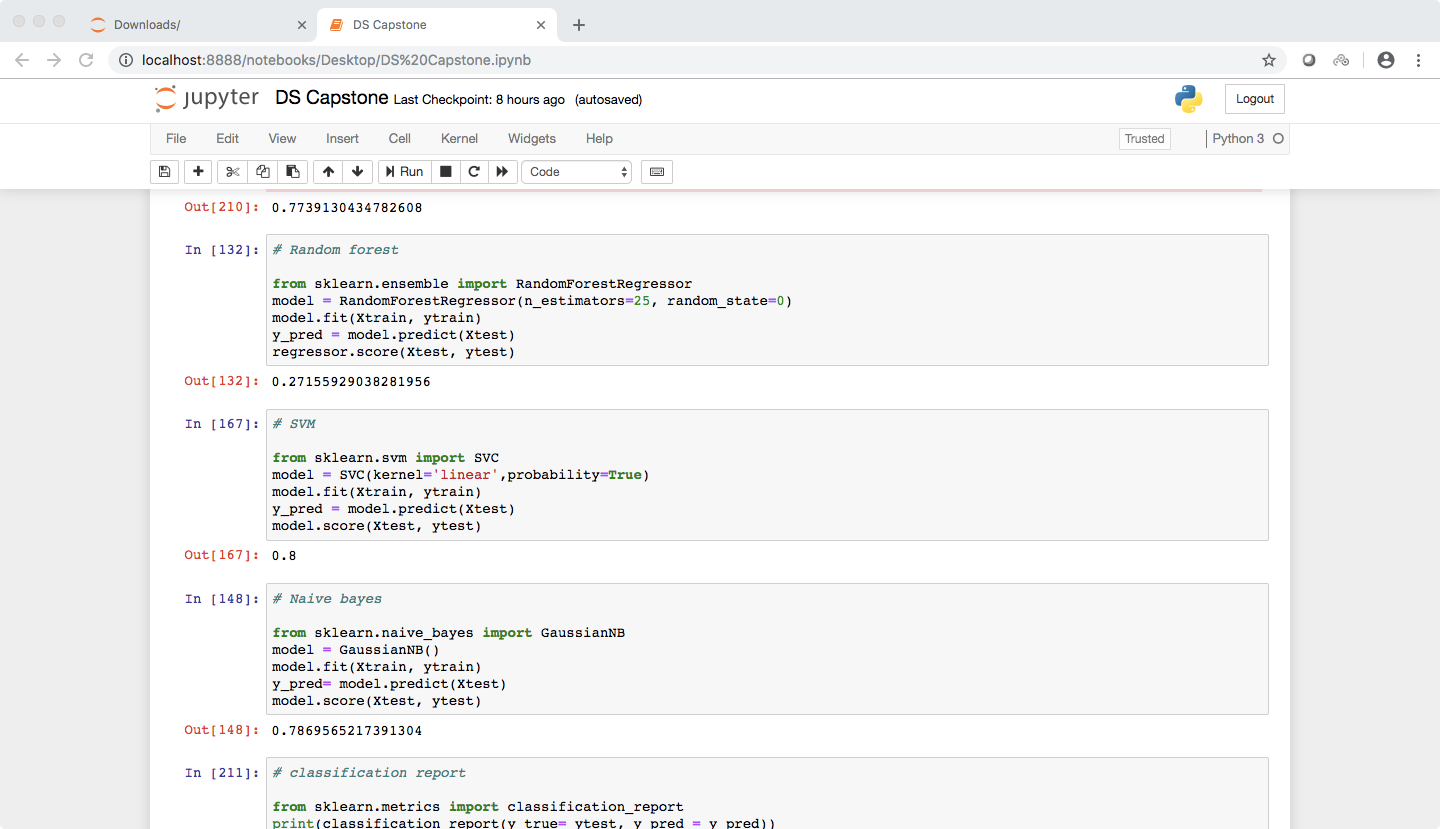


It had a cross validation score of 73.09 %



The third model I used was the SVM.

The accuracy score was 80 %.



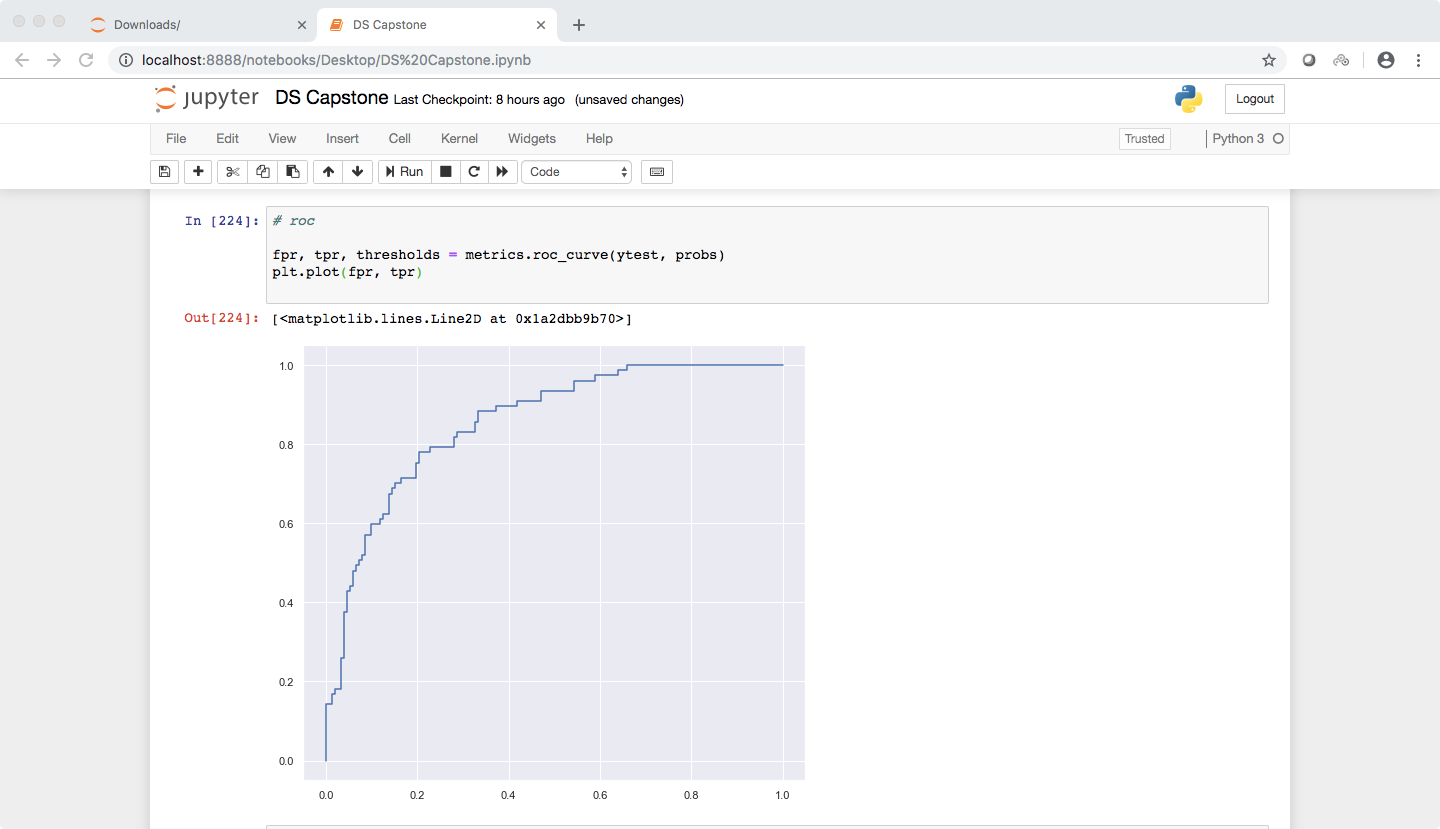
As seen in the classification report below, It had a specificity of 90 % and sensitivity as 60 %

And according to the confusion matrix, it had TP=138 TN=46 FP=15 FN=31

Had an AUC score of 85.57 %

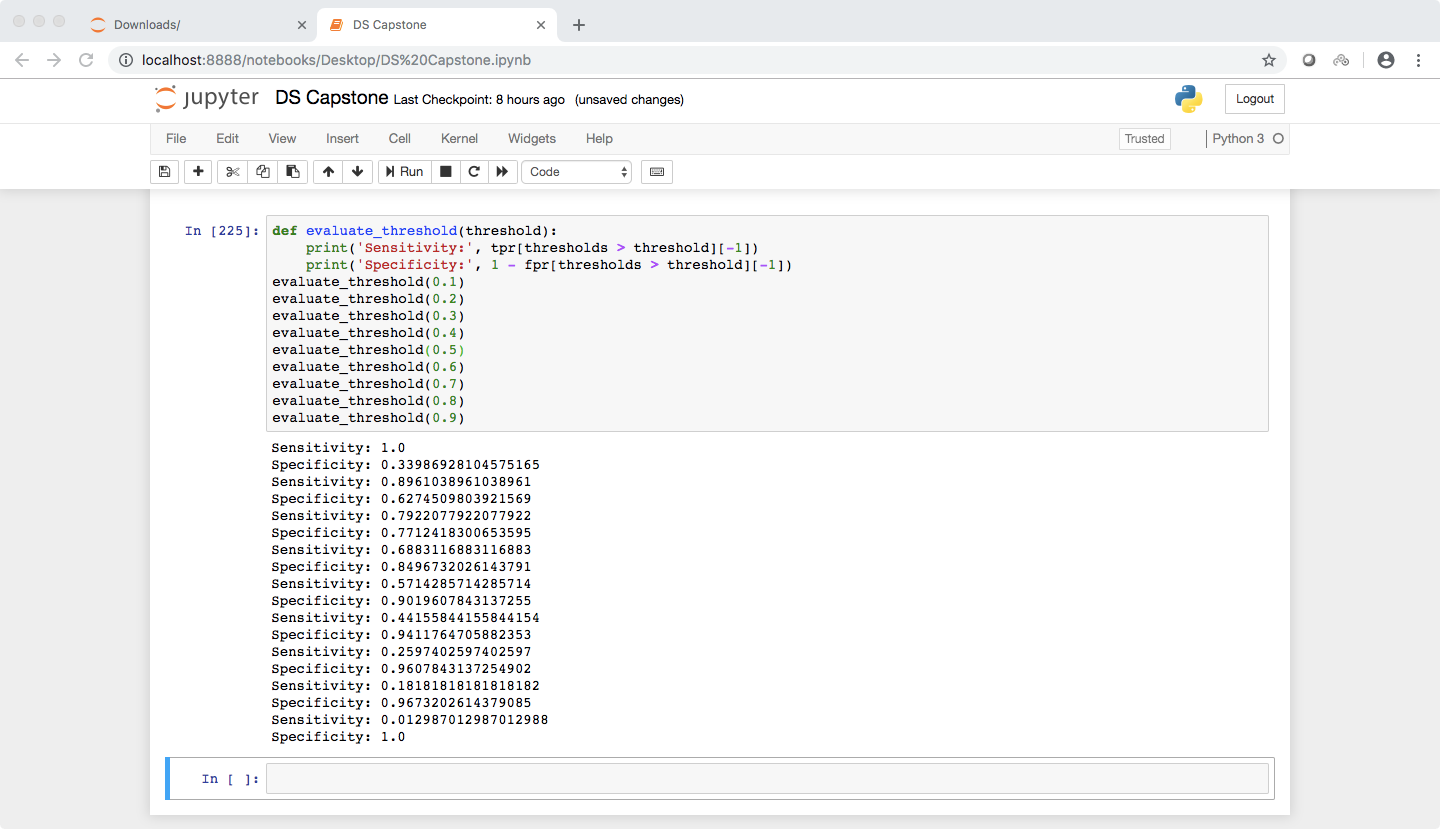


Below is the roc curve

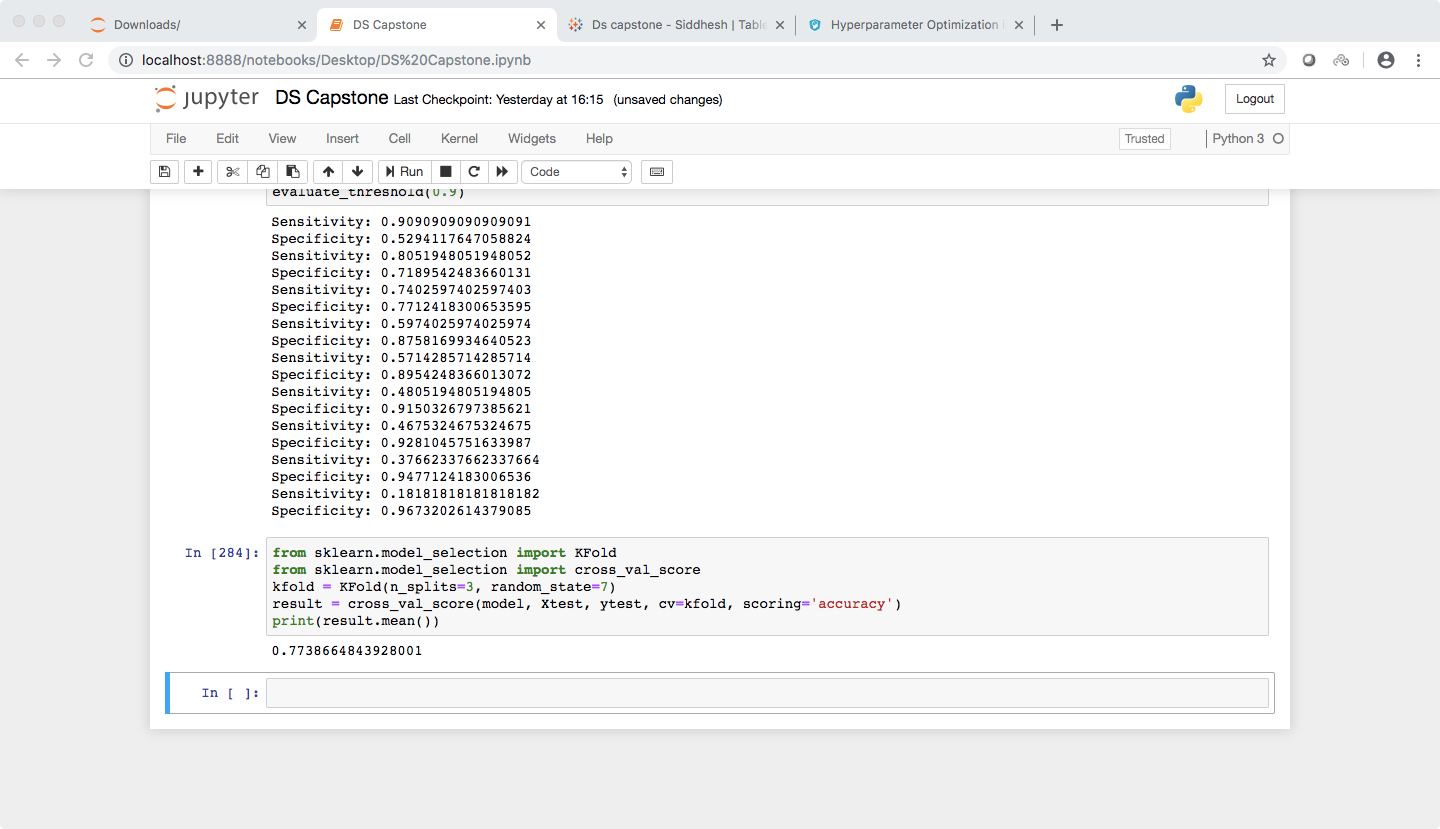


At a threshold of 0.2 it gave specificity = 83 % and sensitivity = 75 %.

So 20 % was the optimal threshold.



It had a cross validation score of 77.38 %



By far this is the best model in my opinion.

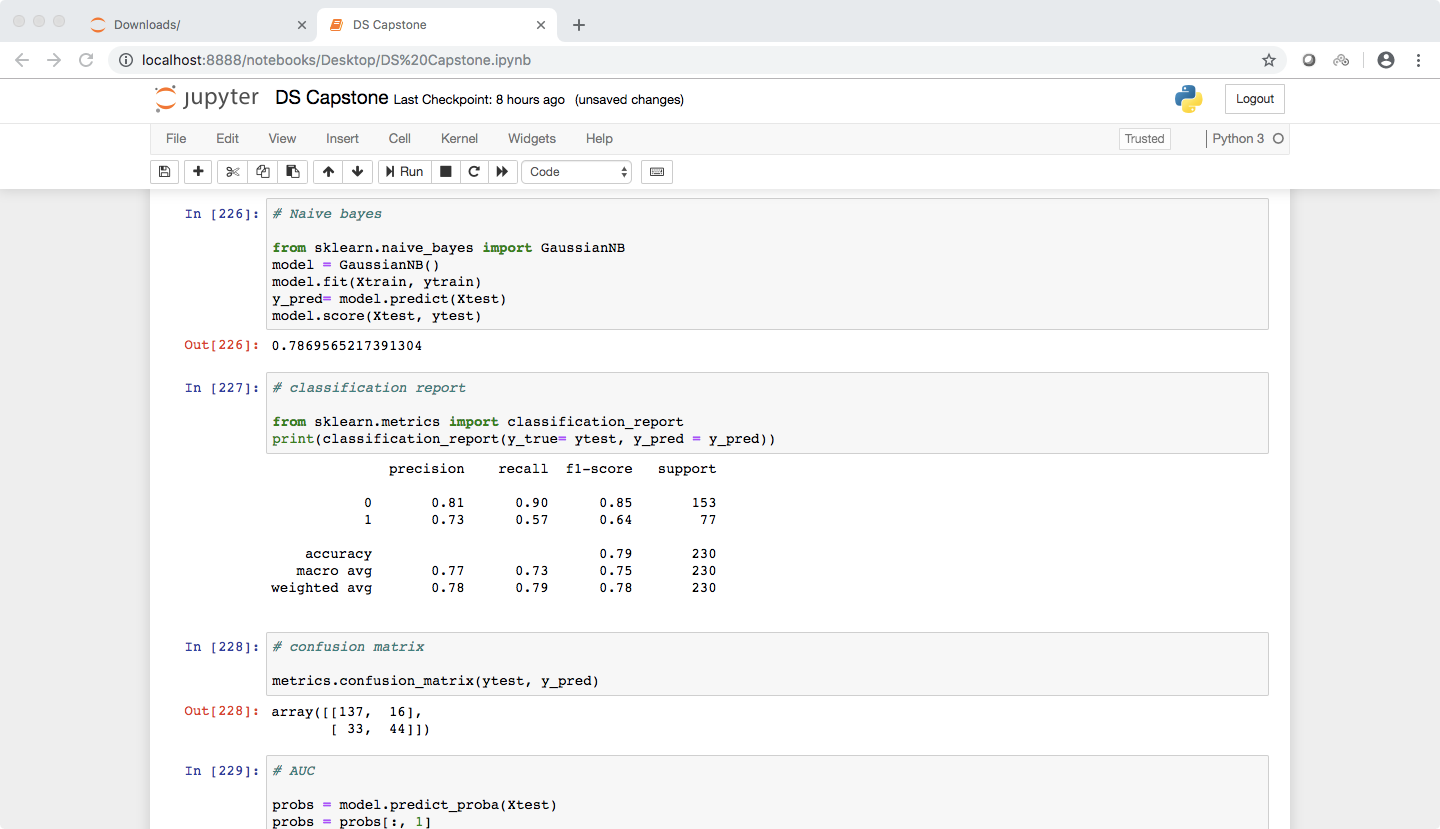
The fourth model I tried was the naive bayes Gaussian model

The accuracy score was 78.69%

As seen in the classification report below, It had a specificity of 90 % and sensitivity as 57 %

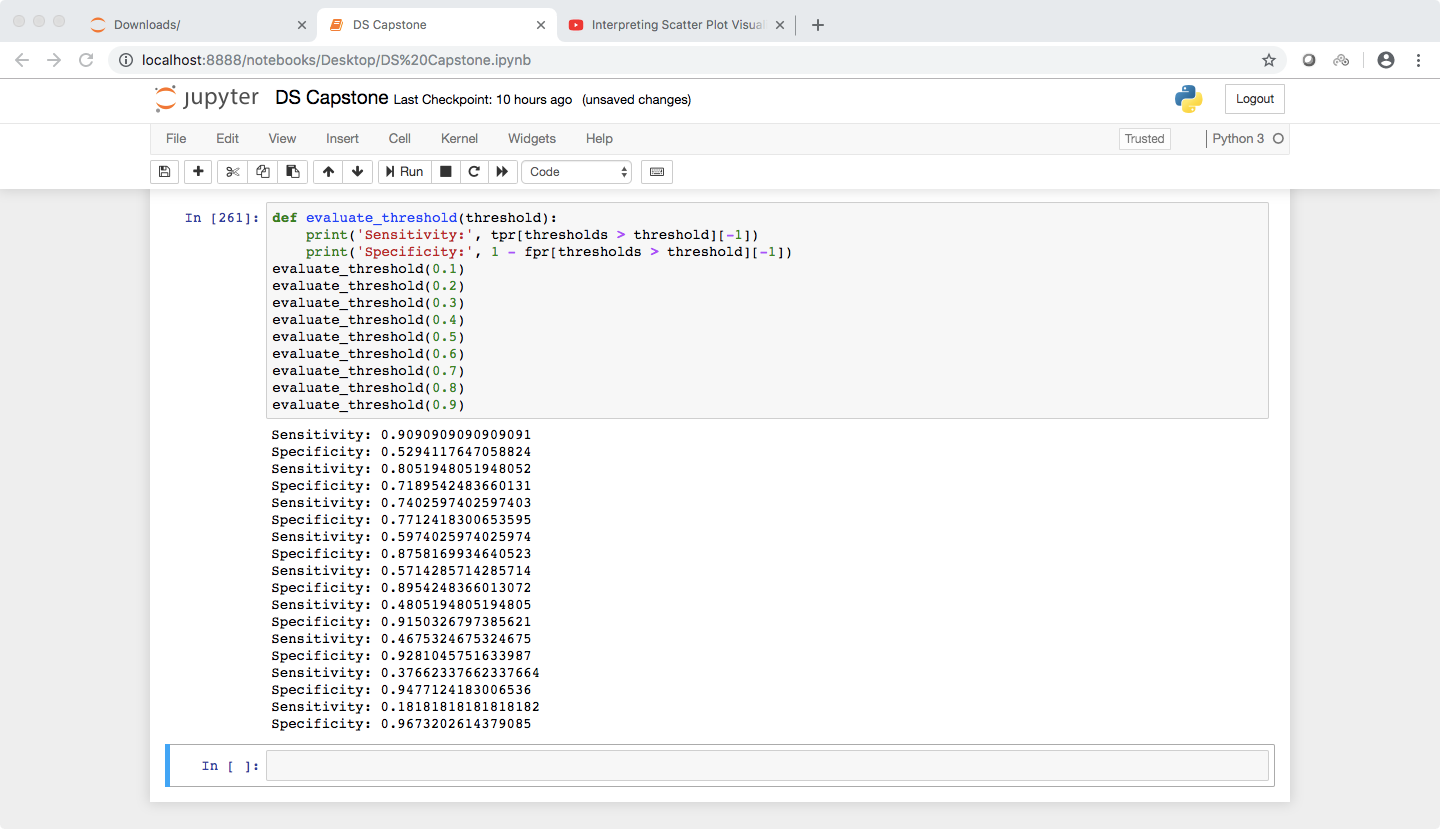
And according to the confusion matrix, it had TP=137 TN=44 FP=16 FN=33

Had an AUC score of 83.57 %

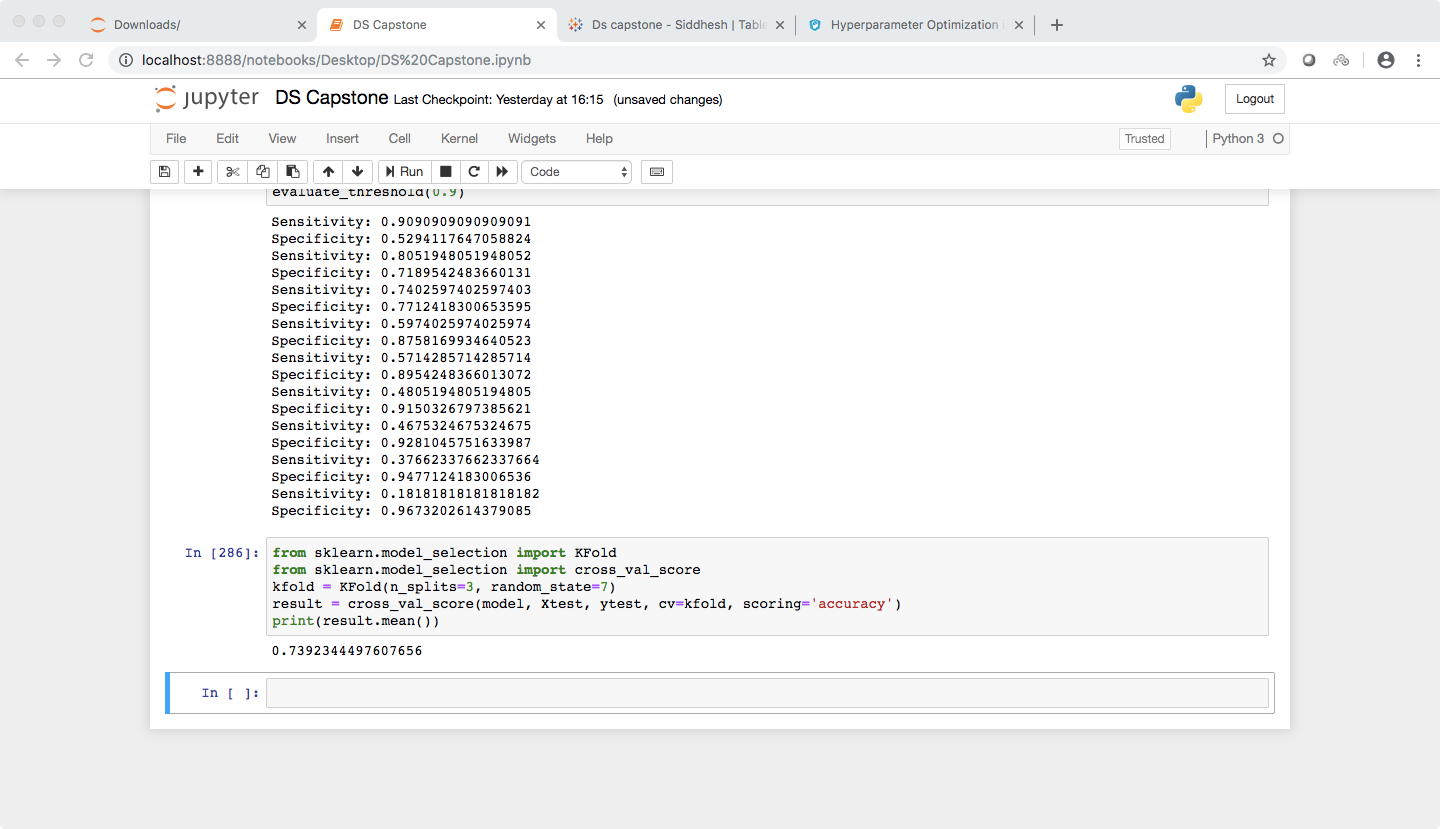


At a threshold of 0.2 it gave specificity = 80 % and sensitivity = 71 %.

So 20 % was the optimal threshold.



It had a cross validation score of 73.92



So as already stated above, I found the SVM to be more efficient than the rest of the models. It had the highest accuracy score, the best (specificity,sensitivity) combination, the best confusion matrix and at 20% threshold it yielded the best levels of specificity and sensitivity.

Here is the link for the tableau part of the project :-

https://public.tableau.com/views/Dscapstone/Dashboard1?:embed=y&:display\_count=yes&publish=yes&:origin=viz\_share\_link