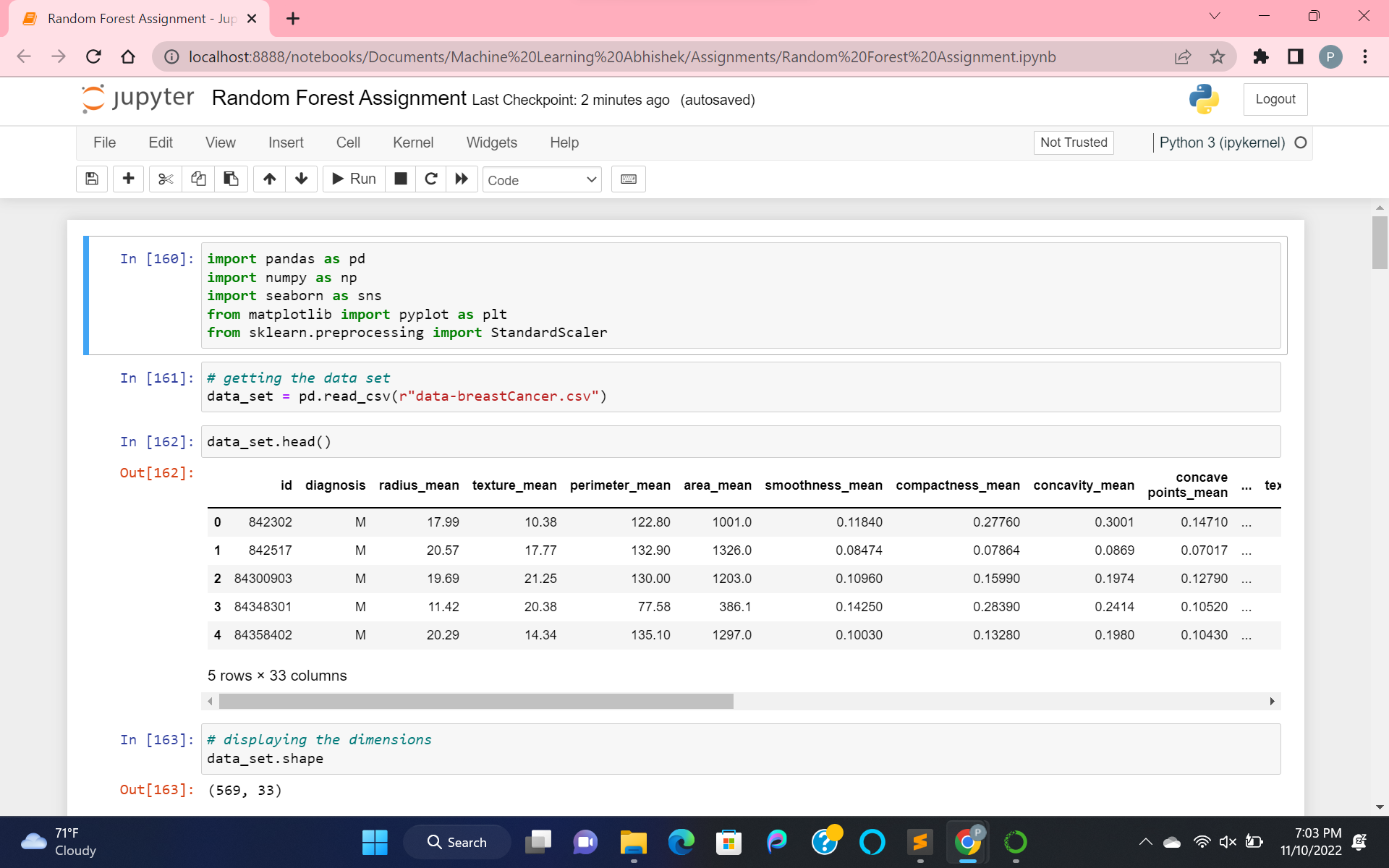
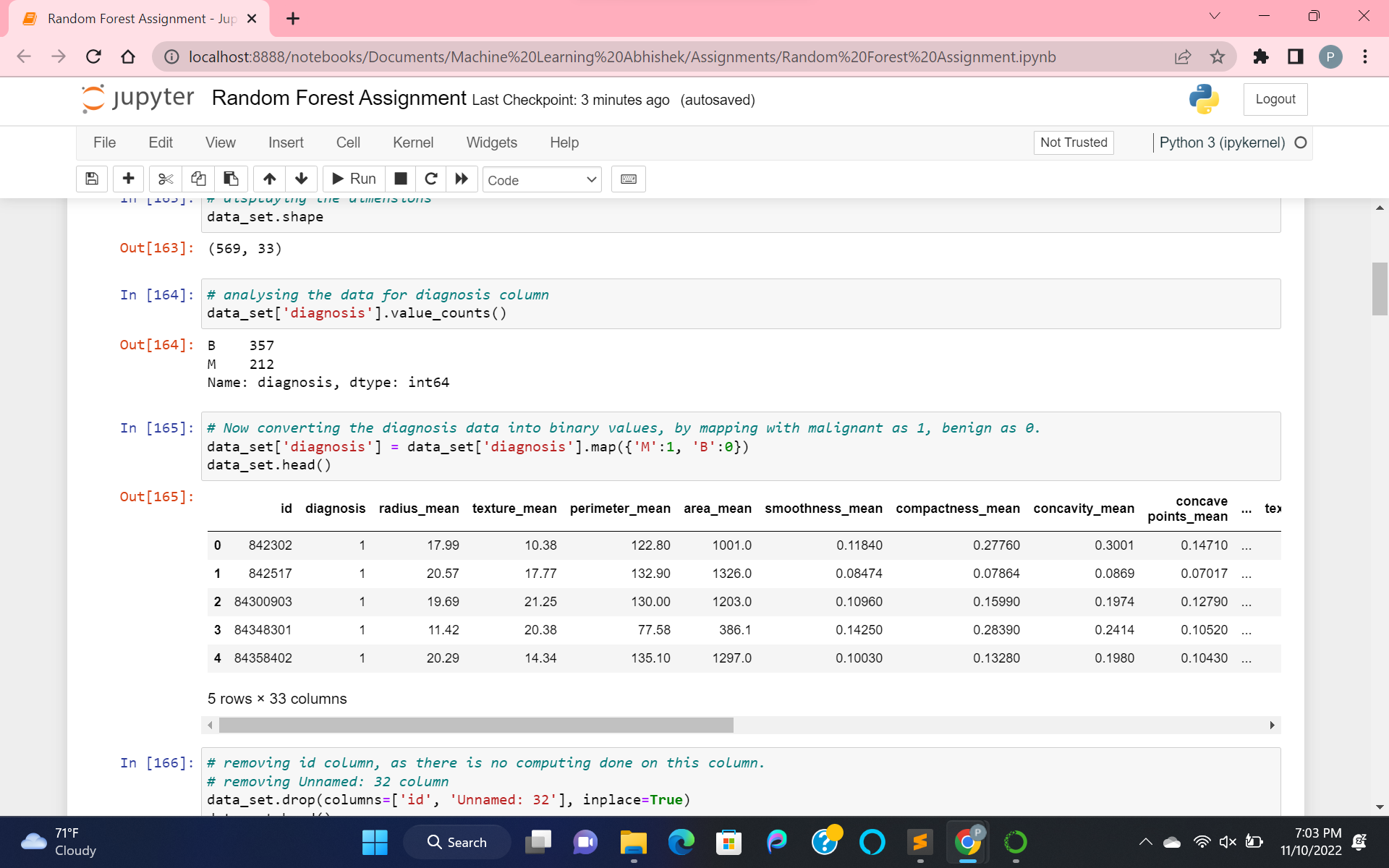
**RANDOM FOREST**

Random Forest is an ensemble technique, which is used to perform both classification and regression tasks with the help of multiple decision trees. The idea is to combine those multiple decision trees in determining the final output.

Given Dataset of Breast cancer wisconsin (diagnostic) dataset

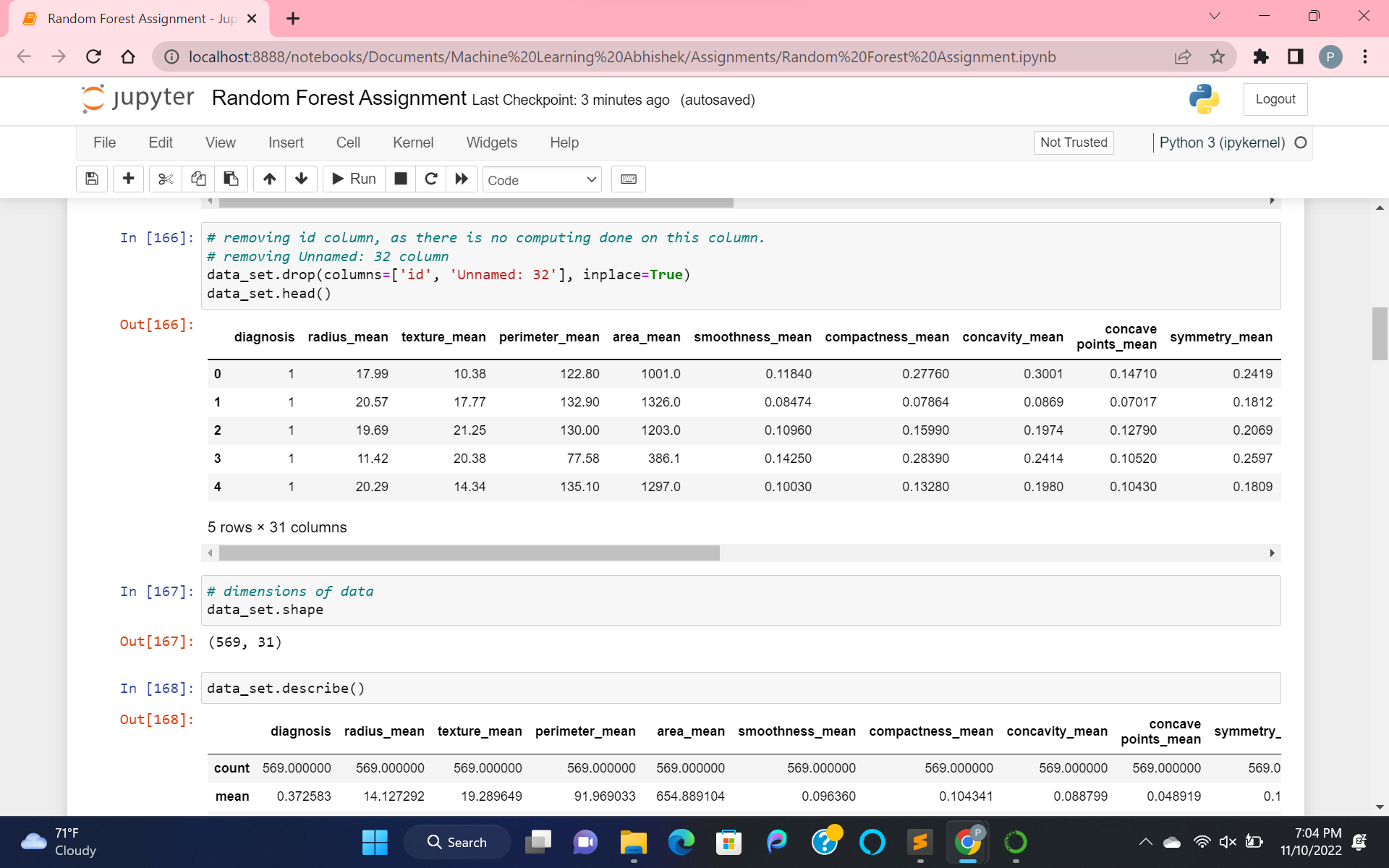


Fetching the data and displaying the dimensions for analyzing the data. There are a total 569 instances and 33 features.

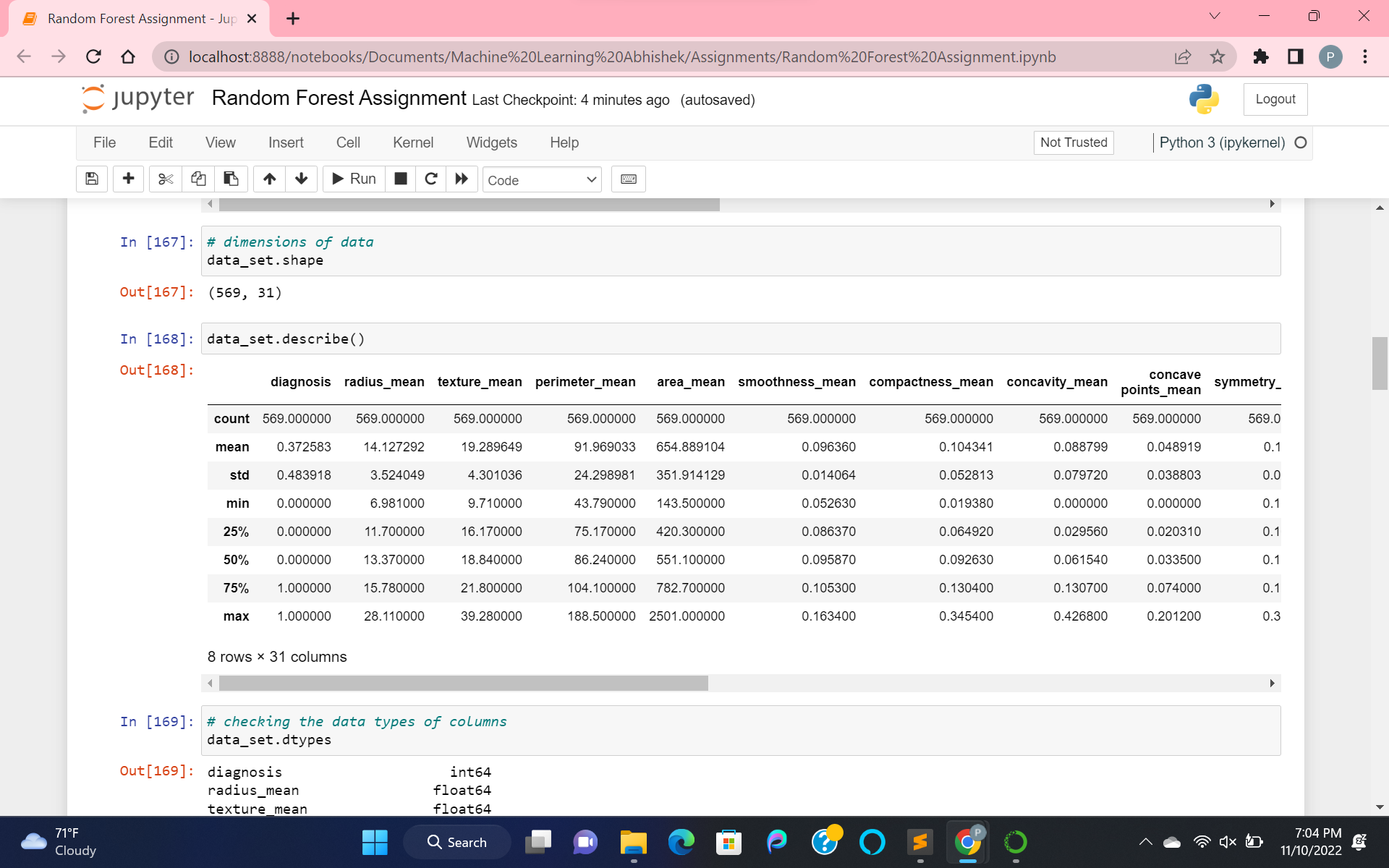


Analyzing the diagnosis column data. It consists of two values: Malignant as “M” with 212 and benign as “B” with 357.

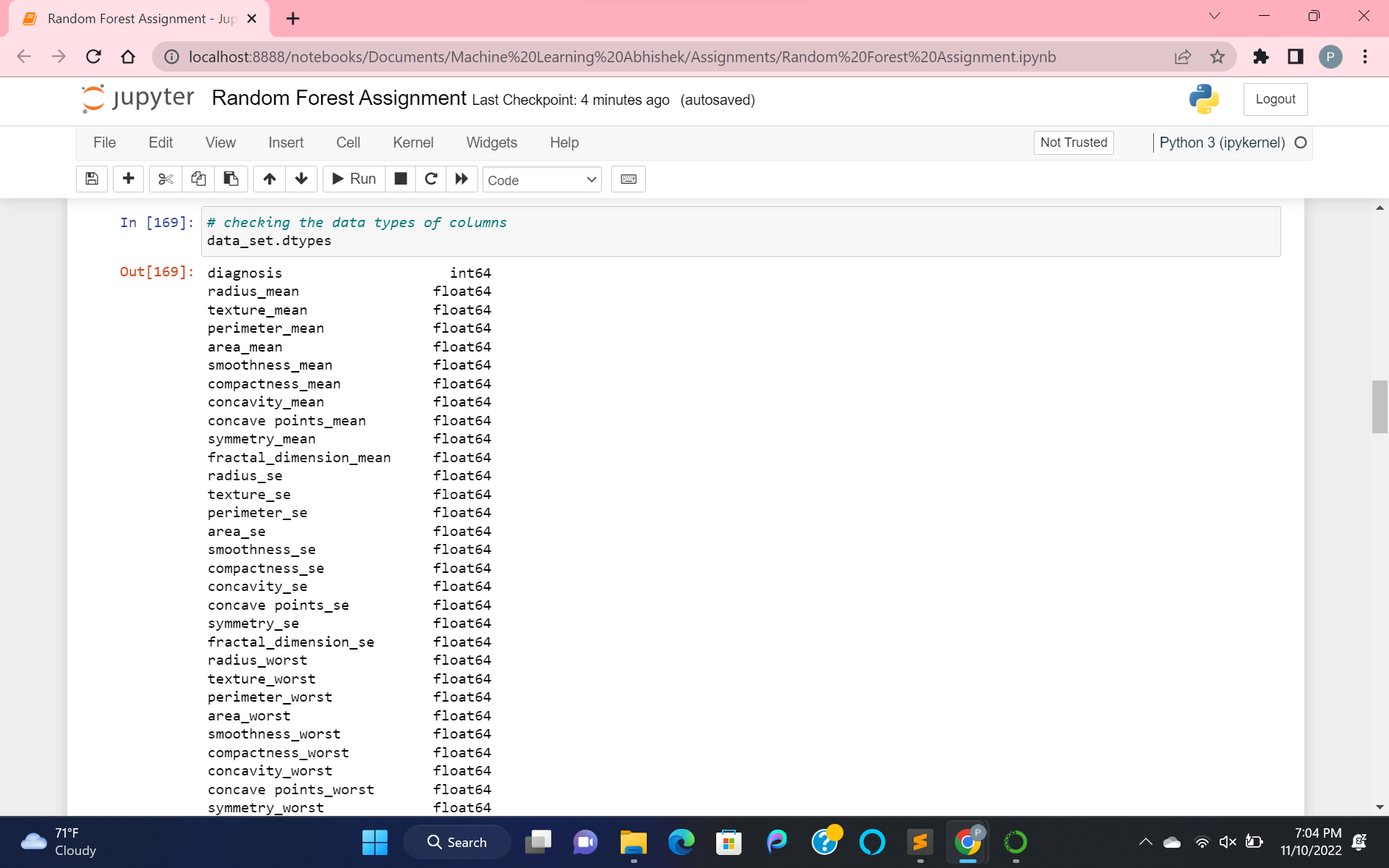
Now Mapping those values labeled as “M” to 1, “B” to 0 for processing of the data.



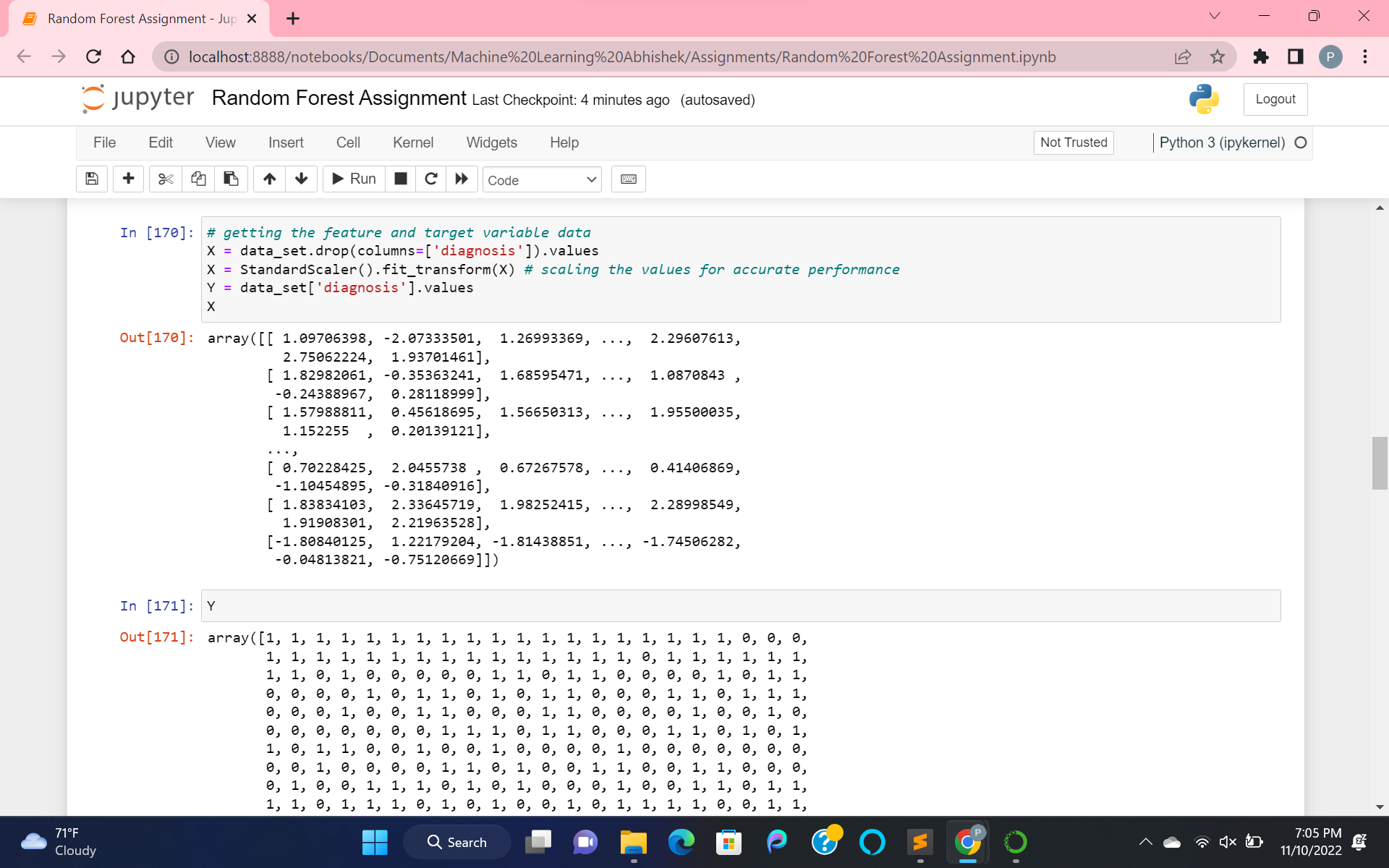
Dropping off the columns that we are not going to use in the model. And rechecking the dimension of the data.



Analyzing the whole dataset by getting the statistical data using the describe() function.



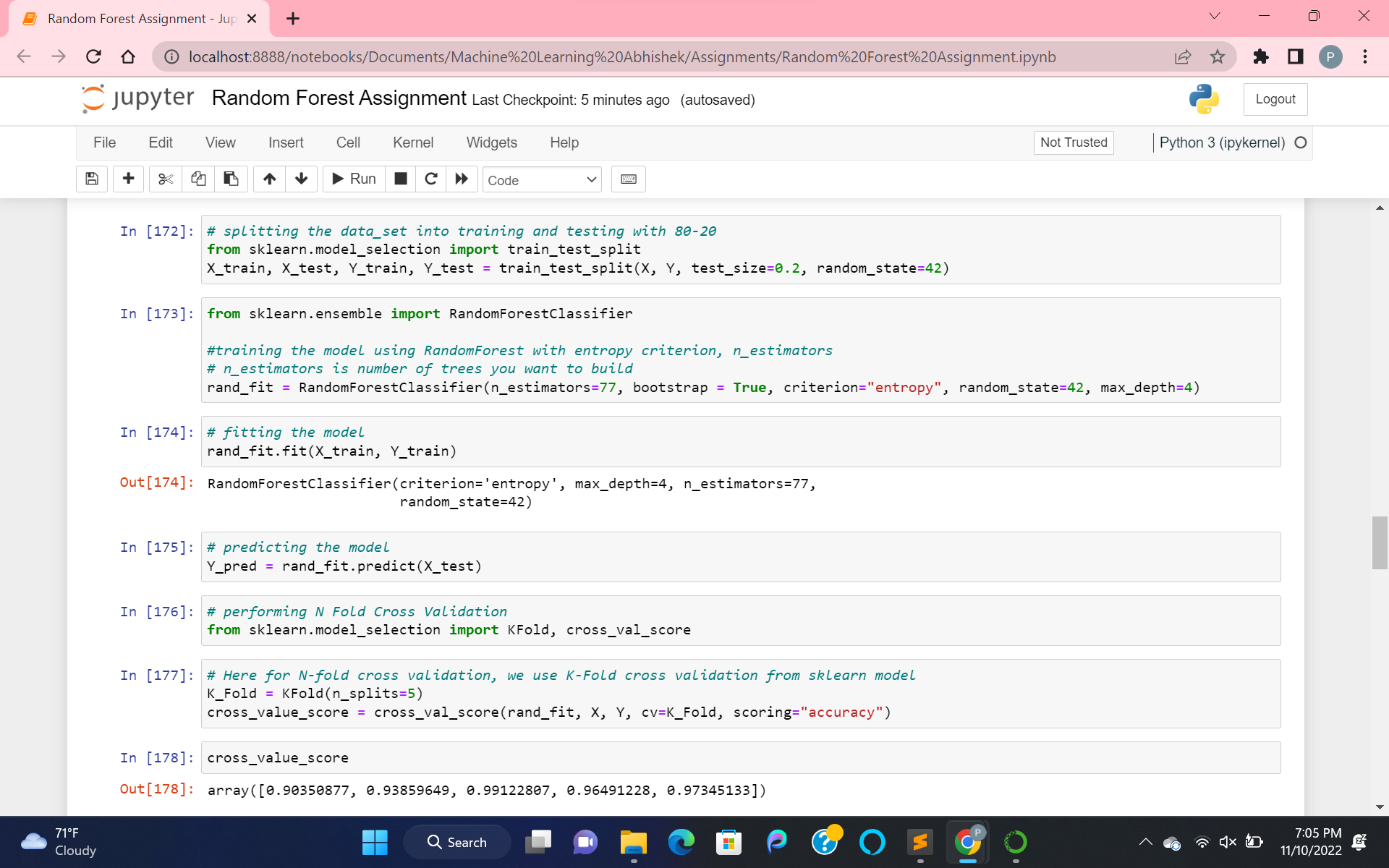
Checking the datatypes of all the features.



Fetching the feature and target variable data.

Taking the diagnosis as the target variable in Y.

Taking the other remaining columns as Independent variables in X. And scaling the independent variables for accurate performance using Standard scaling.



Splitting the data into training and testing data by 80-20 ratio.

Training the model of data using RandomForestClassifier from Sklearn with n\_estimators(number of trees wanted to build), “entropy” as criterion and max\_depth(height of tree). And then fit the model.

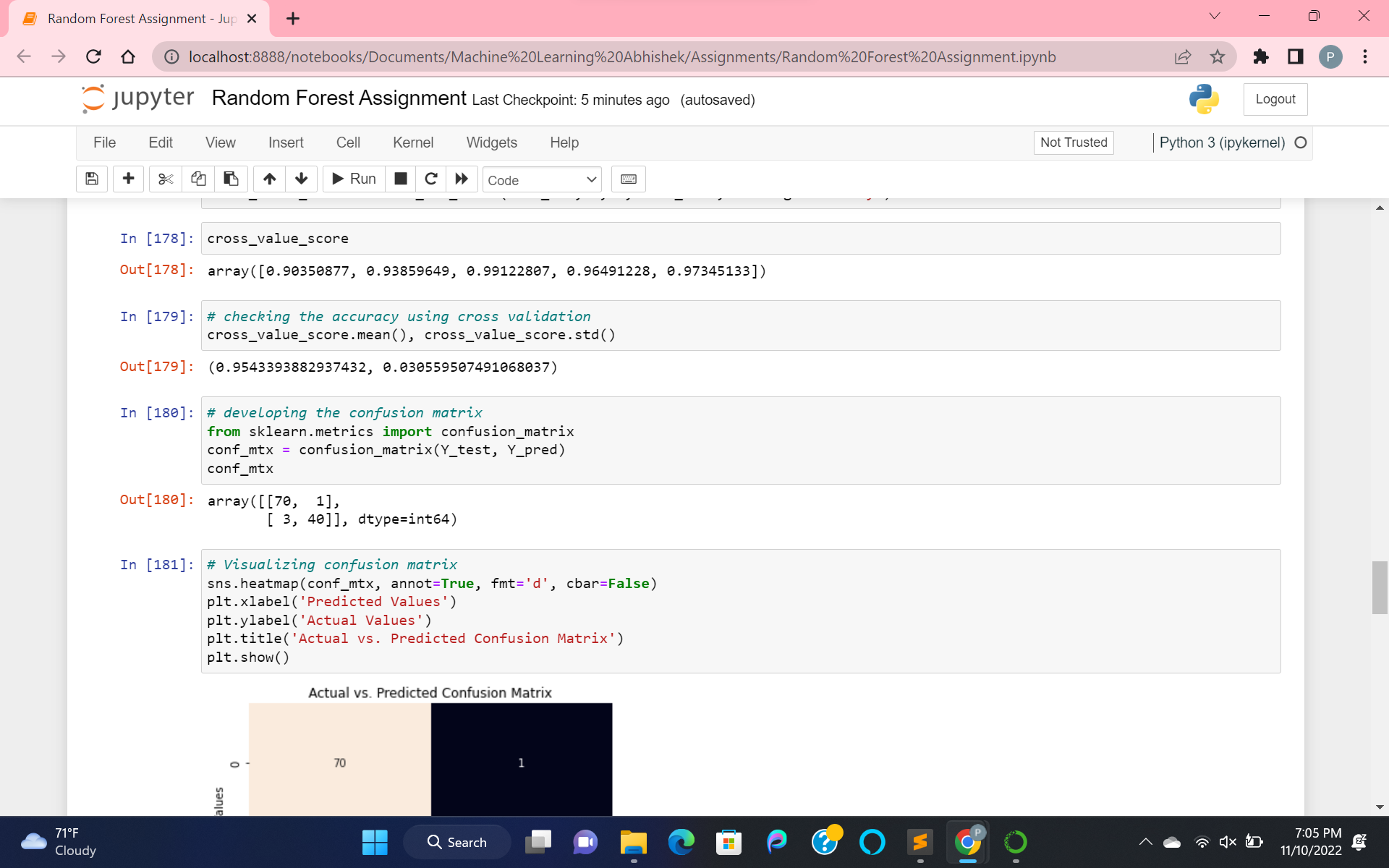
Now, predicting the model by passing testing data.

And then developed the N Fold Cross Validation using K Fold method and cross\_val\_score from sklearn.

Cross Validation is a technique, where we train our model using a subset of the dataset and evaluate using a complementary subset of the dataset.

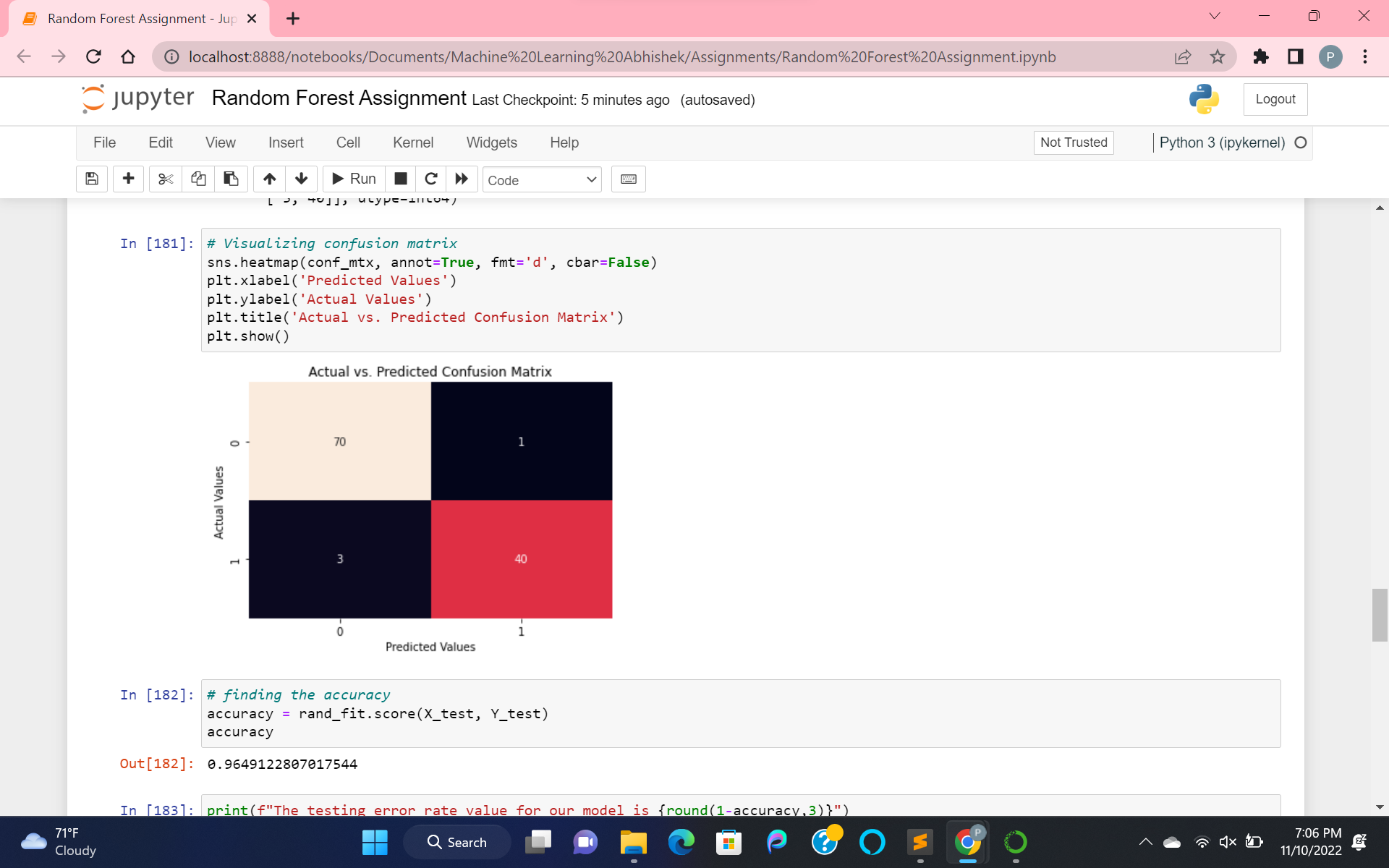
Here n\_splits is the number of splits for the decision tree. And the value I took is 5.

Hence developing the N fold cross validation using RandomForest model and n\_splits and X independent and Y target variables with scoring as accuracy.

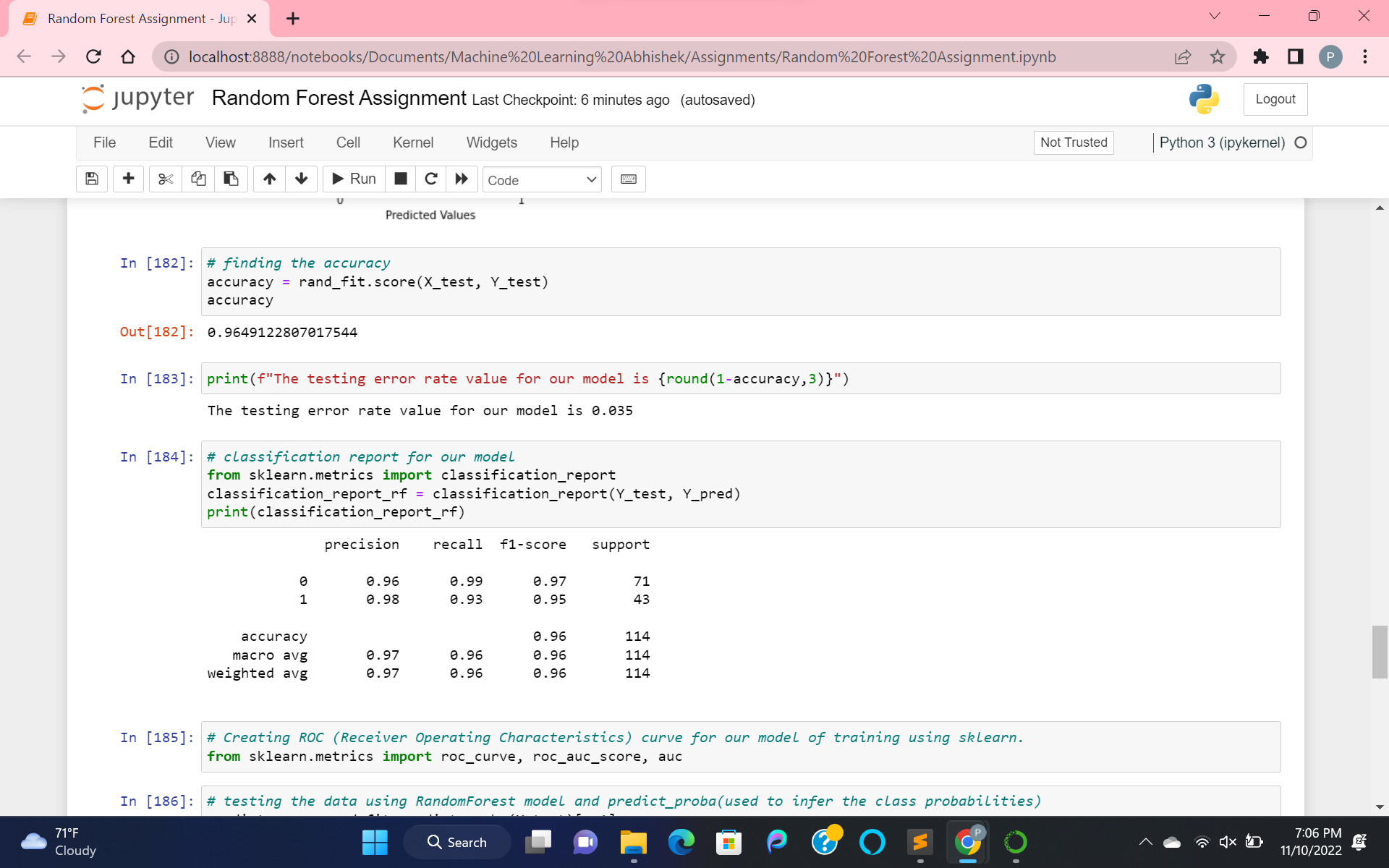


Finding the cross\_value\_score and mean of the cross\_value\_score to measure accuracy.

Then developed a confusion matrix using target test\_data and predicted\_data.



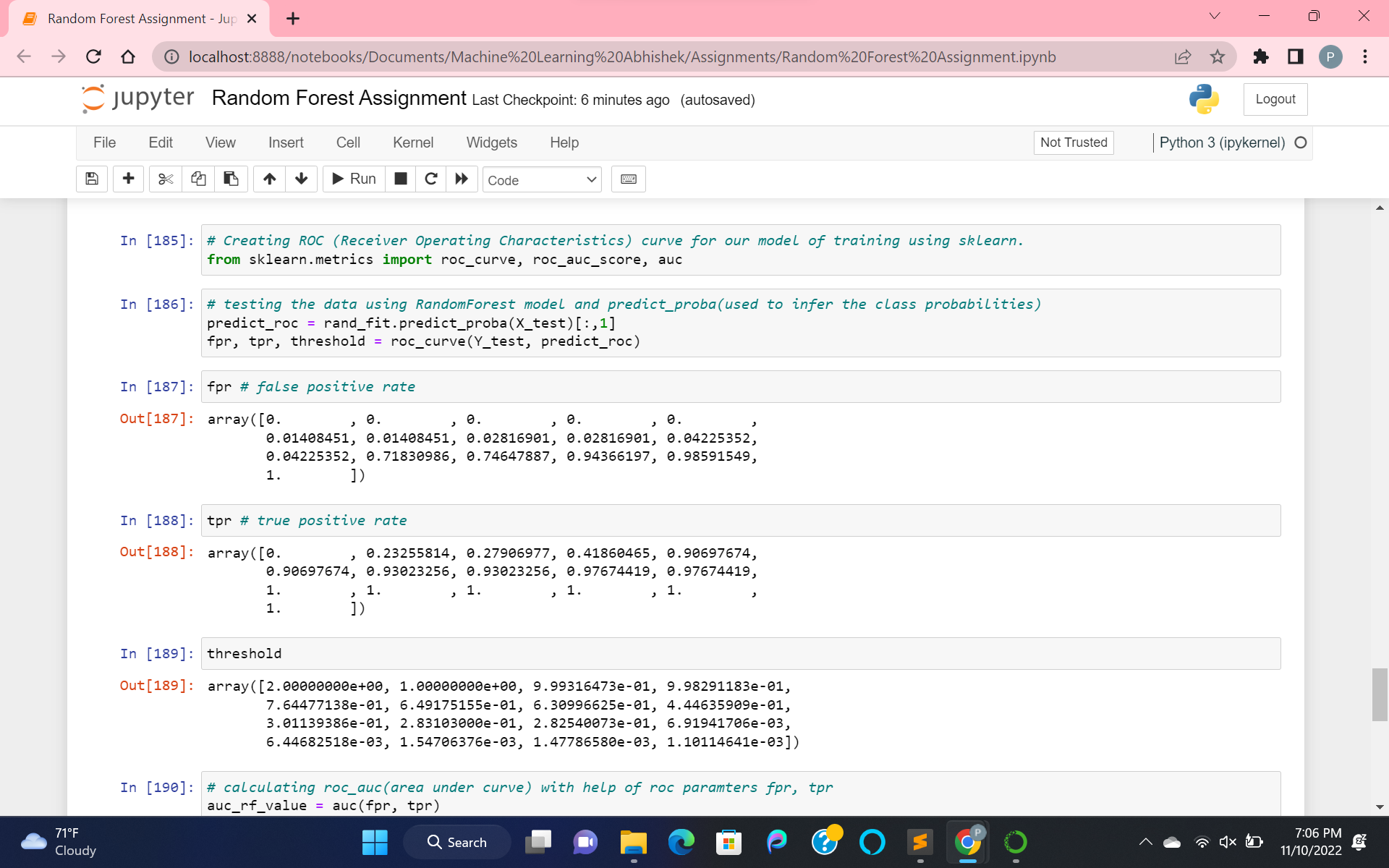
Visualization of confusion matrix with values.



Finding the accuracy of the model and error rate.

Then computed the classification result with targets test and predicted data.

Here class 0 represents Benign, 1 represents Malignant.

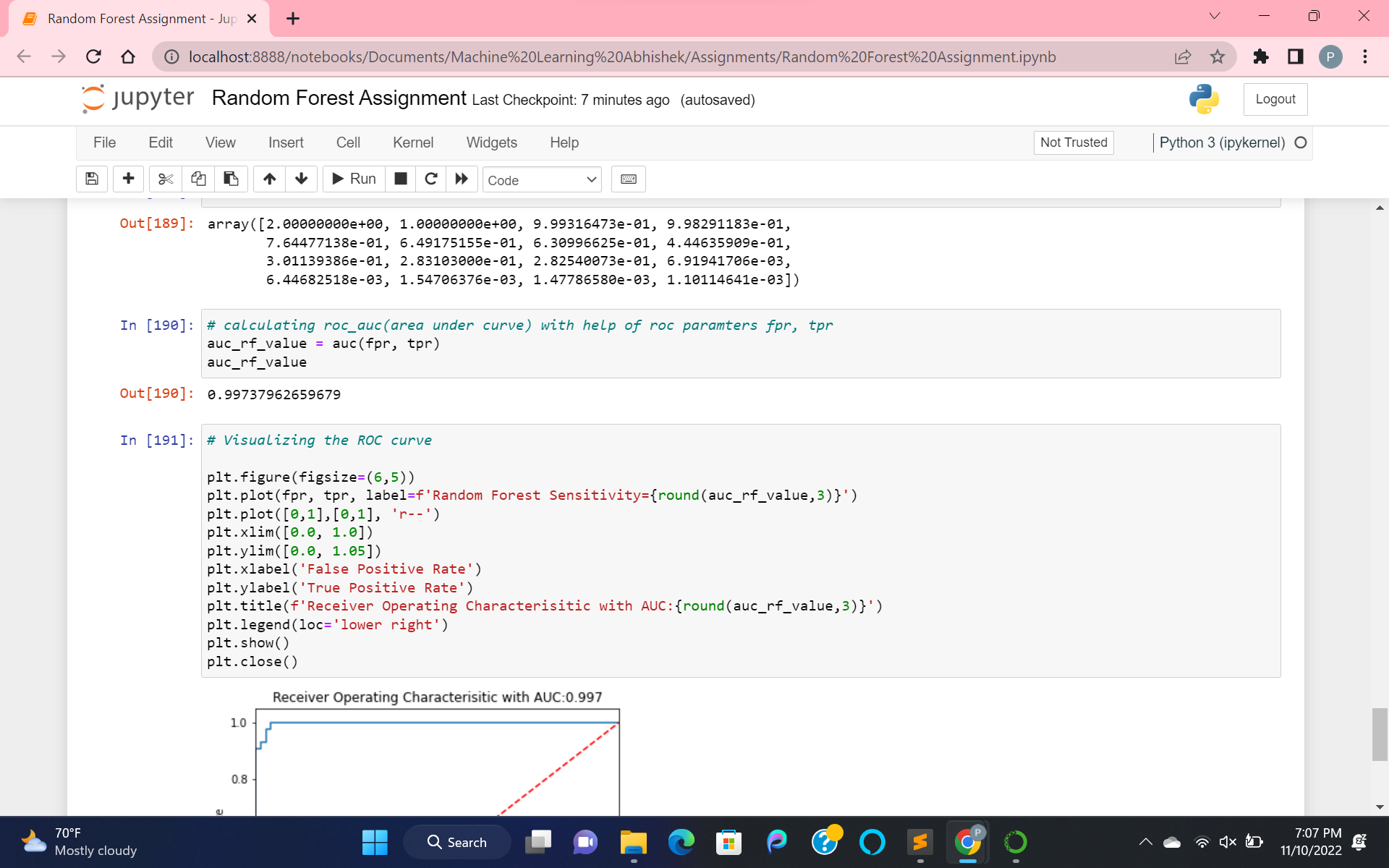


Creating ROC(Receiver Operator Characteristics) with help of training RandomForest model.

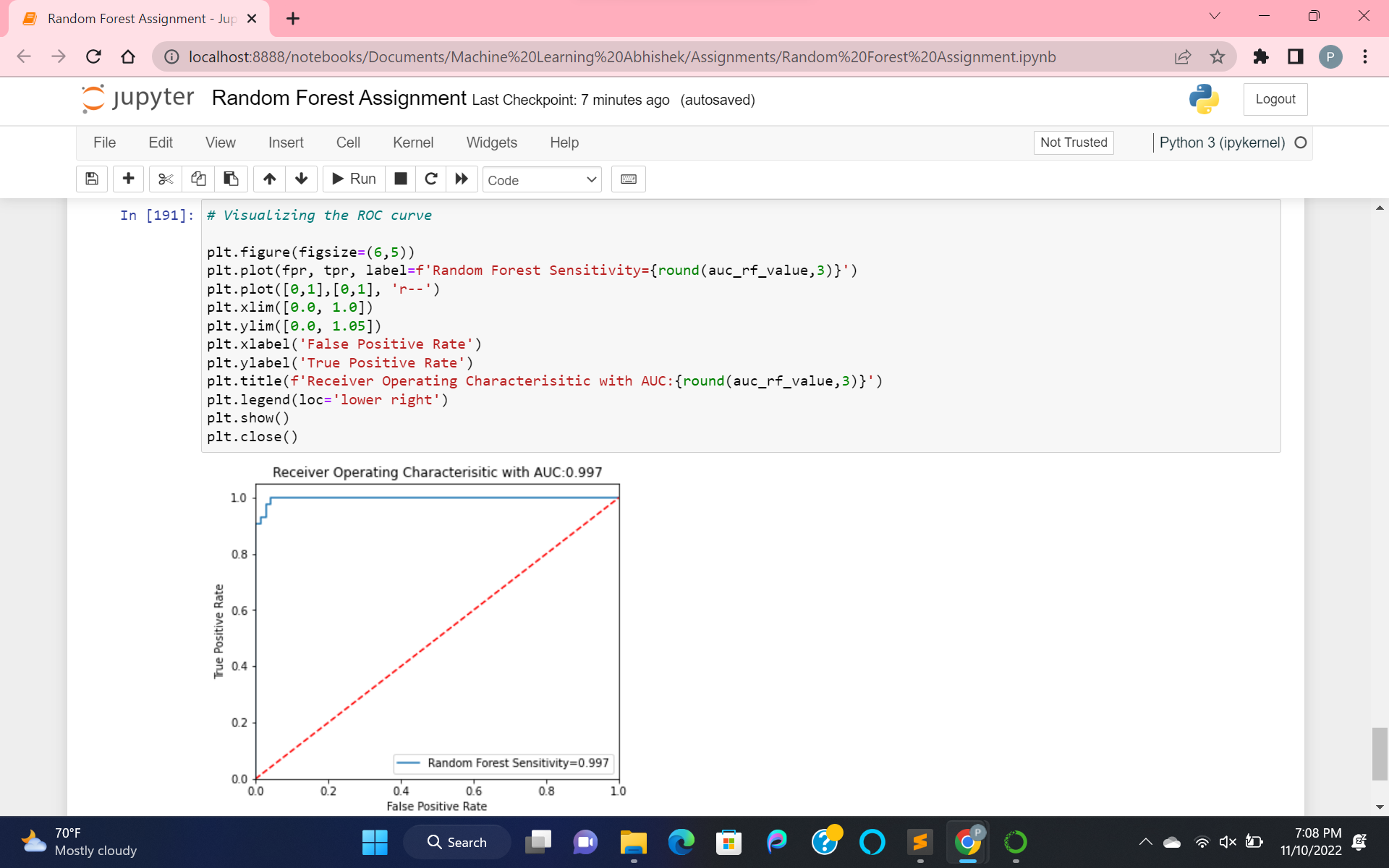
Predicted the data for roc using predict\_probe.

Created roc with help of roc\_curve using target and roc\_predict data with results as FPR(false positive rate), TPR(true positive rate) and threshold.

Here TPR = TP/(TP+FN), FPR = FP/(FP+TN)



Now processing AUC(area under curve) with fpr and tpr as parameters.



Visualizing the ROC curve with AUC. If our curve is located at the top left corner of the plot, that indicates an ideal model. I.e false positive rate of 0 and true positive rate of 1.