Automatic Coronavirus Test Kits

* The Data-set consists of 30000 entries and 28 columns.
* There is only one Categorical column, 26 are numerical columns and one target column.
* We are One-hot encoding the categorical data.
* Splitting the training part i.e. All the columns except the Target column.
* X is the training data and y is the training label.
* Data is well distributed but there are many outliers in the data.
* We cannot remove the outliers so we are using Robust Scaler to reduce the effect of outliers on the model.
* After all the pre-processing of the data, we are passing the data in Random Forest Classifier to get the best features and then we passed the data to 8 different models.
* We have taken three different metrics to evaluate the model that is : Accuracy, precision and recall.
* We have used the scorer function to build a scoring for the cross validation.
* We have used cross validation with personalised scorer to split the data and pass it into the model with defined scoring.
* Then we have calculated the roc curves for each model.
* We have taken the auc\_roc Scores for each model.
* We have plot the combined roc curved derived.
* With the best model we have to hypertune is parameters, so as per the observation seen above we selected Random Forest Classifier to do the classification, so we did hypertuning of the parameters of random forest classifier.
* With the best hyper-parameters, we made a new Random Forest Classifier model.
* This time we included the test data-set so that it be used to evaluate the model.
* Calculated its Auc\_Roc Scores and plot its Roc curve.
* It gave a recall score of 98.9 percent that is almost equal to 1 and a precision score of 94, which is fulfilling our requirement.