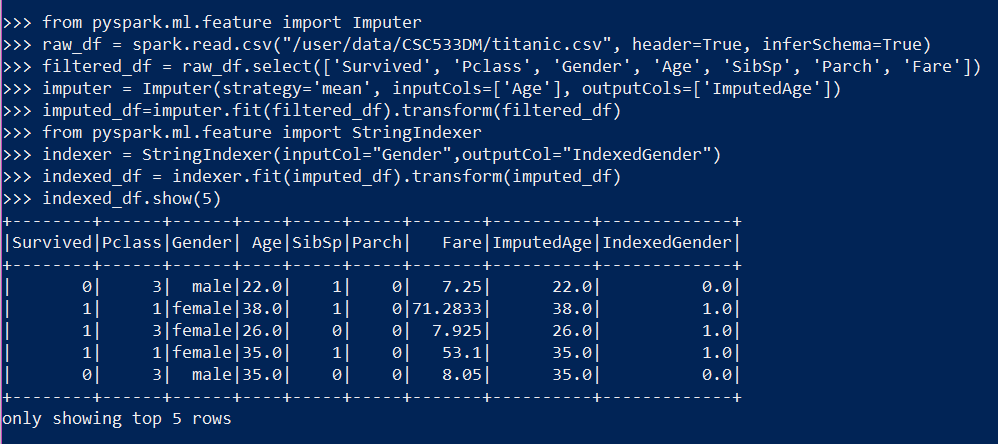
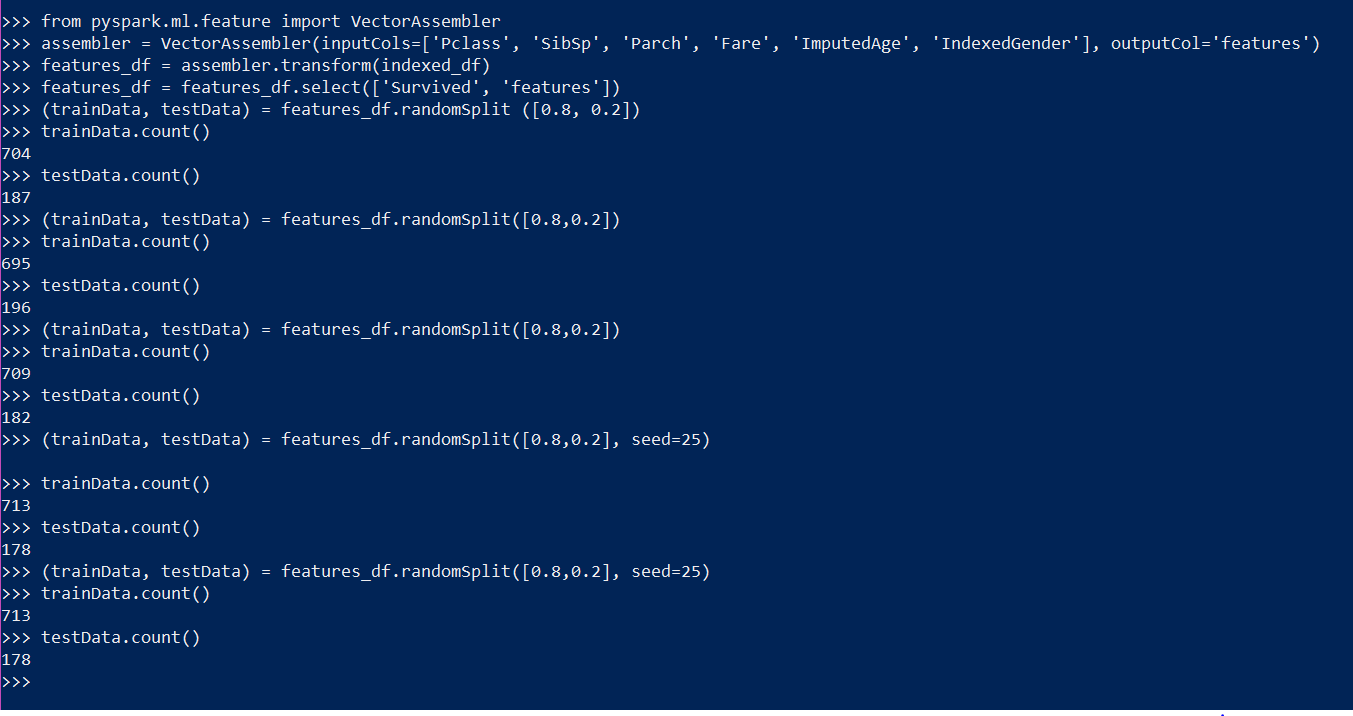
Spark ML: In this assignment, I have tackled missing values with Data Imputation, indexed categorical features by encoding values, used the imputed and encoded values to form a feature vector for training and testing data using randomSplit(). Accordingly, used Random Forest Classifier; It is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Used the area under ROC and area under Precision-Recall curves in a Binary Classification Evaluator class for model evaluation.

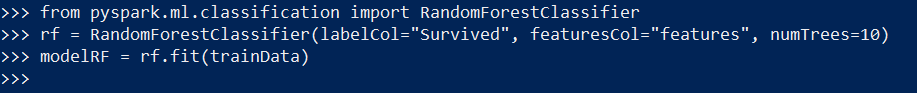
Assignment 1:



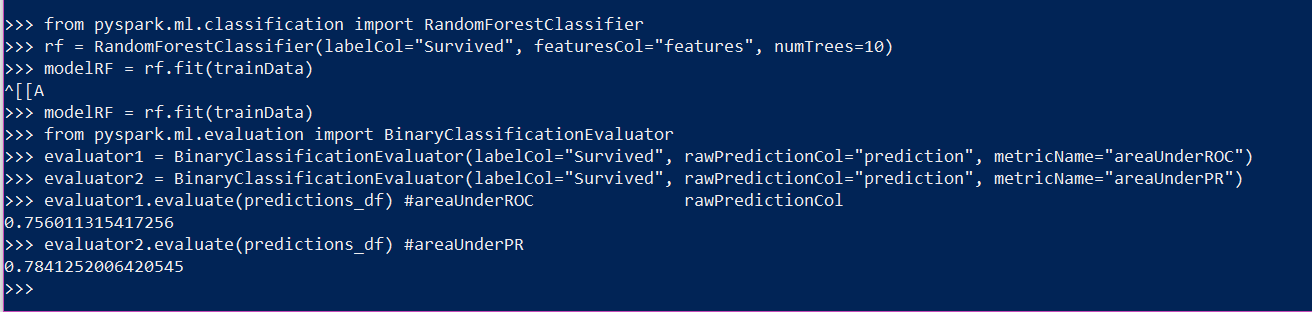
Assignment 2:



Assignment 3:



Assignment 4:



Assignment 5:

1. The BinaryClassificationEvaluator is catered towards models that predict binary labels, i.e 1 or 0, whilst MulticlassClassificationEvaluators are meant for models that tend to predict labels that consist of multiple values. Essentially, they support different metrics for evaluation.

The BinaryClassificationEvaluator supports only Area Under Receiver Operating Characteristic (ROC) whilst the MulticlassClassificationEvaluator supports Accuracy Scores, Weighted Precision and Recall and F1 scores.

2. The Receiver Operating Characteristic (ROC) is a plot generated with True Positive Rate vs False Positive Rate. The ROC curve illustrates the efficiency of a model in predicting good vs. false, inaccurate predictions. If the value of the Area Under the Curve (AUC) is high, that denotes that the performing model is good.

The Precision Recall curve is a plot generated with Precision vs. Recall. The P-R curve illustrates the precision-recall relationship at varying cutoffs. The highlighting difference between them lies in the fact that the number of true negatives are disregarded whilst generating P-R curves.

When we examine the shape of the curves, we interpret that the models perform better when the curve is shaped towards the upper right corner for P-R, and in the case for ROC, the upper left corner.

If there persists a class imbalance problem, P-R curves are preferred. If there is no class imbalance problem, ROC Curves are used.