**Software Engineering and Data Science, SEIS 763: Machine Learning**

**Assignment #9 (100 points)**

**Due Date: November 29th**

**PRAGYA VERMA (101276204)**

1. Train a single decision tree from the training set to predict the breast cancer types

in the testing set.

**Ans:**

*# Importing libraries*  
**import** pandas **as** pd  
**from** sklearn.tree **import** DecisionTreeClassifier  
**from** sklearn.tree **import** plot\_tree  
**from** sklearn.metrics **import** classification\_report,confusion\_matrix

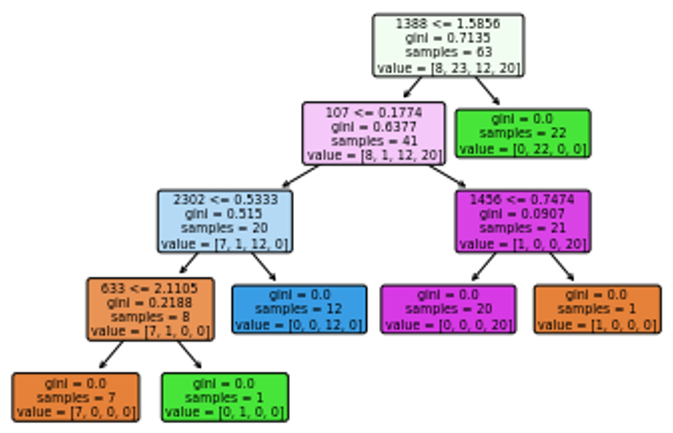
**import** pandas **as** pd  
file\_path = **r**"C:\PRAGYA\UST\COURSES\SEM 4\Machine Learning\Assignments\HW9\ML\_HW\_Data\_CancerGene.xlsx"  
  
*# Importing the data from different sheets of the Excel file*Cancer\_Gene\_train\_X = pd.read\_excel(file\_path, sheet\_name=0, header=None)  
Cancer\_Gene\_train\_labels = pd.read\_excel(file\_path, sheet\_name=1, header=None)  
Cancer\_Gene\_test\_X = pd.read\_excel(file\_path, sheet\_name=2, header=None)  
Cancer\_Gene\_test\_labels = pd.read\_excel(file\_path, sheet\_name=3, header=None)

*# Veryfing test\_lables*  
Cancer\_Gene\_test\_labels.head()

*# Python interprets the class 'NA' as if it were a NaN. So change the class label from 'NA' to 'NAA'.*  
Cancer\_Gene\_test\_labels.loc[Cancer\_Gene\_test\_labels[0].isnull(), 0] = 'NAA'

*# Training and fitting the DecisionTreeClassifier on the Training set*  
DT\_Model = DecisionTreeClassifier(criterion = 'gini' )  
DT\_Model.fit(Cancer\_Gene\_train\_X, Cancer\_Gene\_train\_labels )

*# Plotting the decision tree*  
DecTree = plot\_tree(decision\_tree=DT\_Model, feature\_names = Cancer\_Gene\_train\_X.columns, filled = True , precision = 4, rounded = True)

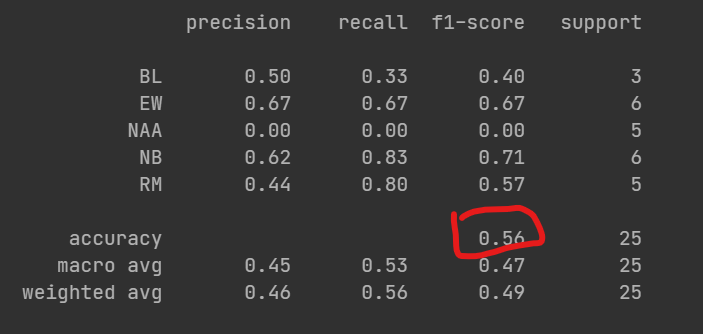


2. From test data, what is the precision / recall / F-measure for \*\*EACH\*\* class?

**Ans:**

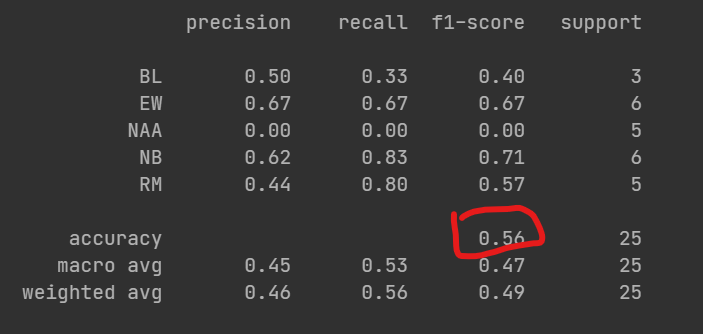
*# Prediction of Test Data*  
Predict\_test = DT\_Model.predict(Cancer\_Gene\_test\_X)

*# Classification Report for Test Data*  
print(classification\_report(Cancer\_Gene\_test\_labels, Predict\_test))

****

3. Do you see anything special from your observation in Question 2?

**Ans**: This model is classifying different types of data and correctly predicted 56% of test data since its overall accuracy is 0.56 which can be considered good but it could be improved. Across different classes such as BL, EW, NAA, NB, RM, the model's performance varies. It is performing very well on class label EW with precision of 0.67 and f1-score of 0.67 but performing very badly on class labelled NAA with the precision and f1-score are zero, the model did not predict anything and model is not recognizing any true positives for this class which is problematic. Also, there is class imbalance in the dataset with varying numbers of instances for each class (e.g., 3 for BL and 6 for NB could be the reason for discrepancy and affecting models’ performance.

****