

CLASSIFICATION METRICS

You have developed a decision tree model for classifying genetic samples. This week, you will learn about methods that measure the effectiveness of classification models. You will use the techniques to evaluate the effectiveness of your decision tree model.

Complete the following tasks:

1. Compute the accuracy for each feature (individual genetic mutation). Which feature yields the highest accuracy? Show a table of the top 10 features (ranked by accuracy) with their corresponding accuracy values.
2. In this exercise, you will use a decision tree to classify a collection of samples and compute metrics to evaluate the effectiveness of the decision tree classifier. Previously, you constructed a decision tree classifier using the “**TP – FP**” quantity to select the best features for classifying samples. This resulted in a set of decision rules. Use the decision rules to *classify all samples* provided in the data file. Calculate the following metrics to characterize the resulting set of classifications of *all samples*:
 - TP, FP, TN, FN
 - Accuracy
 - Sensitivity
 - Specificity
 - Precision
 - Miss rate
 - False discovery rate
 - False omission rate

To do this, construct a confusion matrix to characterize how your decision rules classify the set of all samples. Then, use the information in the confusion matrix to compute the metrics listed above.

Be prepared to provide a demo in class next week, including the following:

- a table of the top 10 features ranked by accuracy (with corresponding accuracy values),
- the result of classifying all samples (including a confusion matrix),
- the values of the metrics listed above and
- a discussion and interpretation of the preceding items.

NOTES:

1. *You are no longer permitted to use Excel for data mining activities. You must write a computer program for all subsequent activities in the course.*