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 methods to evaluate the effectiveness of your decision tree model.

**Materials**:

* Dr. Welch’s Introduction to Classification Metrics: <https://youtu.be/jmrTGqIUNEY>
* <https://en.wikipedia.org/wiki/Sensitivity_and_specificity>

C**oncepts to learn from the materials:**

Be able to *(1) define, (2) list alternate names,* and *(3) write the formulas* for the following classification metrics:

* Accuracy
* Sensitivity
* Specificity
* Precision
* Miss rate
* False discovery rate
* False omission rate

**Quiz**:

After learning the concepts listed above, complete the BlackBoard quiz no later than Monday September 27, 9:39 am. The quiz will cover your understanding of the *concepts to learn from the materials* (see above). The quiz may include multiple choice, true-false, fill-in-the-blank, and/or matching questions.

Data Mining Activity: (*to be started after you complete the quiz*)

* **Due date**: no later than Thursday Sept. 30, 11:59 pm
* Submit results, report and computer program by email to [welch@ohio.edu](mailto:welch@ohio.edu)

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 you will compute metrics to evaluate the effectiveness of the decision tree classifier. Previously, you constructed a decision tree classifier by using the quantity “**TP** – **FP**”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 the set of all samples is classified by your decision rules. Then, use the information in the confusion matrix to compute the metrics listed above.

Submit an email to [welch@ohio.edu](mailto:welch@ohio.edu) that contains a brief report, including the following:

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

Additionally, attach the computer program that you developed for this activity and the output of your program (either a screenshot(s) or a file).

**NOTE***: you must develop your own computer program to accomplish this assignment. You ARE NOT permitted to use pre-existing programs for building decision trees, computing evaluation metrics, or any other component of this project.*

**NOTE**: I may respond to your email submissions with questions about your methods, results, and/or interpretation. Please respond promptly to my questions.