DECISION TREES and DECISION RULES

This week, you will develop a *multifactor* genetic test to determine if an individual may develop cancer. You will build upon the decision tree that you constructed last week.

Recall that last week you performed the tasks listed below:

1. Compute the quantity “**TP** – **FP**” for each feature (genetic mutation). Which feature yields the maximum value of “**TP** – **FP**”? Show the table of the top 10 features (ranked by their corresponding values of “**TP** – **FP**”) and their corresponding “**TP** – **FP**” values.
2. Select the most useful feature, **F**, for correctly classifying samples (i.e., the genetic mutation **F** that yields the maximal value for the quantity “**TP** - **FP**”).
3. Divide the samples into two groups by using the feature **F** to classify each sample as either
   1. group-A: samples that have mutation **F**

or

* 1. group-B: samples that do not have mutation **F**

1. Make a confusion matrix to represent the classification of all samples with genetic mutation **F**.

This week, you should complete the following tasks:

1. Repeat steps 1 & 2 for group-A, resulting in the selection of a feature **A** *that* *best classifies the samples* *in group-A*. Show the table of the *top 10 features* *for classifying samples in group-A* (ranked by their corresponding values of “**TP** – **FP**” *for group-A*) and their corresponding “**TP** – **FP**” values *for group-A*. Make a confusion matrix to represent the *classification of samples in group-A* with genetic mutation **A**.
2. Repeat steps 1 & 2 for group-B, resulting in the selection of a feature **B** *that best classifies the samples in group-B*. Show the table of the *top 10 features* *for classifying samples in group-B* (ranked by their corresponding values of “**TP** – **FP**” *for group-B*) and their corresponding “**TP** – **FP**” values *for group-B*. Make a confusion matrix to represent the *classification of samples in group-B* with genetic mutation **B**.
3. The result is a ‘decision tree’ that contains the following:
4. a root node **R**
5. two children of **R**:

* node **A**
* node **B**

1. two children of **A**:

* node **A1**
* node **A2**

1. two children of **B**:
   * node **B1**

* node **B2**

1. Draw your resulting decision tree (*you may manually draw the resulting decision tree; you are not required to develop a computer program that displays your tree*). Label each internal (non-leaf) node (i.e., nodes **R**, **A**,and **B**) with the genetic mutation used to divide samples at that node. Label the arcs coming from a node with each of the possible values (‘0’ or ‘1’) of the genetic mutation that is represented by the node. Label each leaf node (i.e., nodes **A1**, **A2**,**B1**, and**B2**) with the appropriate class label, signifying the specific class that would be assigned by the tree to all samples at that node (see step 9 for a description of the classification procedure).
2. Your ‘decision tree’ can be used to classify each sample **S** by using the following classification rules:

If **S** has mutation **F** then

if **S** has mutation **A**

then classify **S** as **C**

else classify **S** as **NC**

else

if **S** has mutation **B**

then classify **S** as **C**

else classify **S** as **NC**

1. Use your decision tree to classify the following samples: *C1, C10, C30, NC5, and NC15*. For each of these samples, report which classification your decision tree would assign to the sample. How many times did your decision tree make the correct classification?

Be prepared to provide a demo at the start of the next class session, including the following:

* an annotated drawing of your decision tree (see item 8 above),
* tables of the top 10 features for the internal nodes of the decision tree, ranked by their corresponding values of “**TP** – **FP**” (see items 5 & 6 above),
* confusion matrices for the internal nodes of the decision tree (see items 5 & 6 above),
* your *specific* classification rules (see item 9 above),
* the results of classifying specific samples (see item 10 above), and
* a discussion and interpretation of your results.

**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.*
2. *You must develop your own computer program to accomplish this assignment. You ARE NOT permitted to use pre-existing programs for building decision trees or any other component of this project.*
3. *All assignments submitted for this course should represent your thinking and effort and be prepared entirely by you. Using generative AI at any stage of your work in this course constitutes academic dishonesty and violates course policy and the Ohio University Student Code of Conduct.*