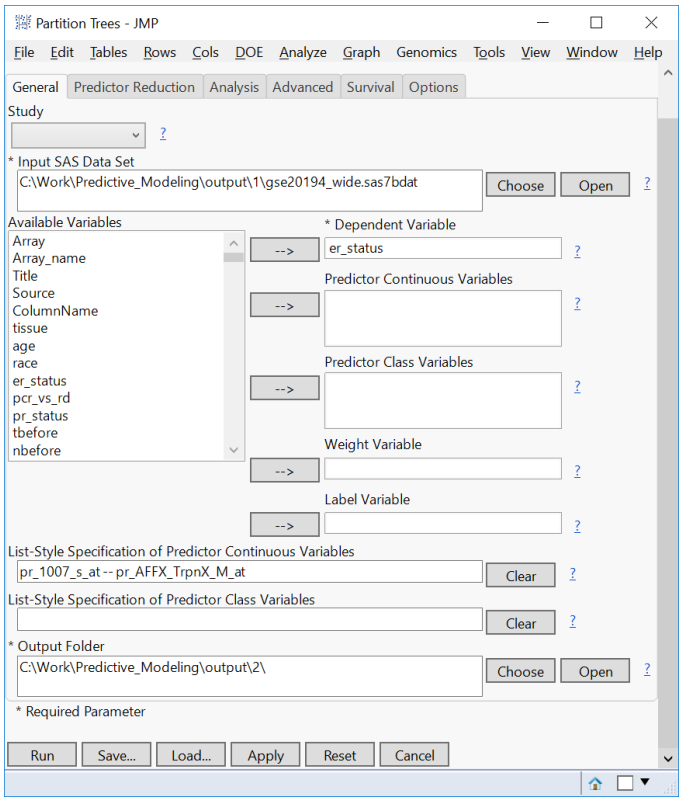
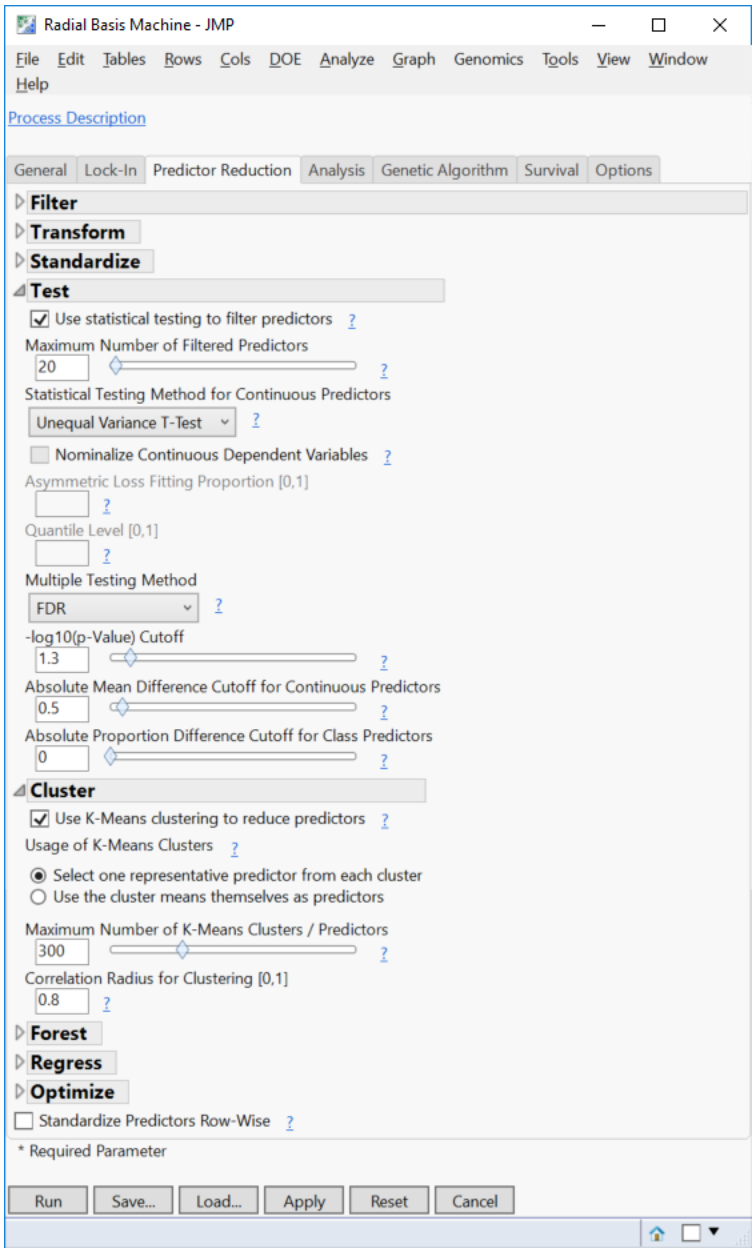
Whole Model Fit

Through previous analyses surrounding the GSE20194 data set downloaded from GEO, we have created models and tested them on **Test Data Sets** to evaluate their fit. In the *Cross Validation Model Comparison Step-Guide*, we created and evaluated a **Partition Trees** and **Radial Basis Machine** model. We then determined our data had an adequate number of samples in the *Learning Curves Step-Guide*. Cross validation and test set model comparison have shown that the **Partition Trees** model is the best fitting model. The last step in the predictive modeling workflow is to fit the whole model to the **Training Data Set** and to assess its effectiveness.

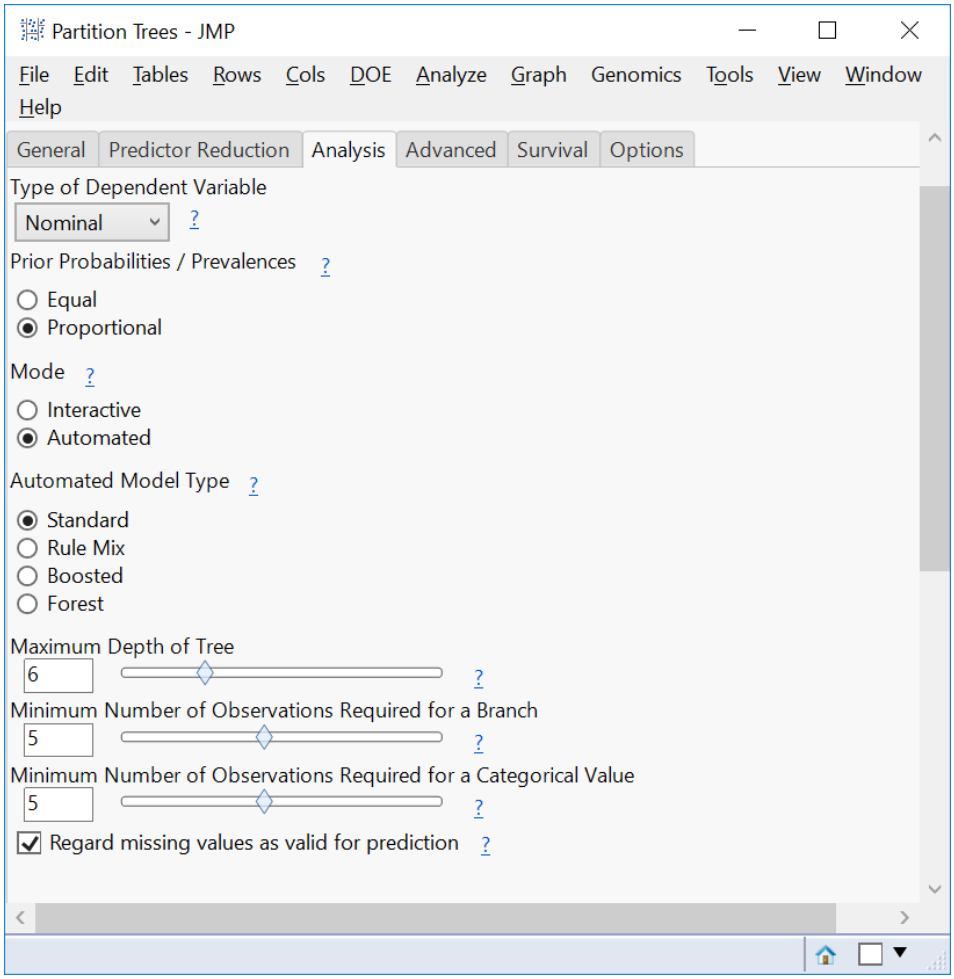
1. From the **Genomics Starter** load the **Partition Tree** model settings. Select **File > Load Life Sciences Setting** from the toolbar. Then find the settings folder created in the *Cross Validation Model Comparison Step-Guide.* The tabs should appear as below.
    - The input data is the ***gse20194_wide.sas7bdat*** set created in the *Importing and Transposing a Data Set Step-Guide*.
    - Note that these settings can be created again by selecting **Predictive Modeling > Main Methods > Partition Trees** from the **Genomics Starter** menu.
    - For further information on the model settings see the *Cross Validation Model Comparison Step-Guide.*



General tab.



Predictor Reduction tab.

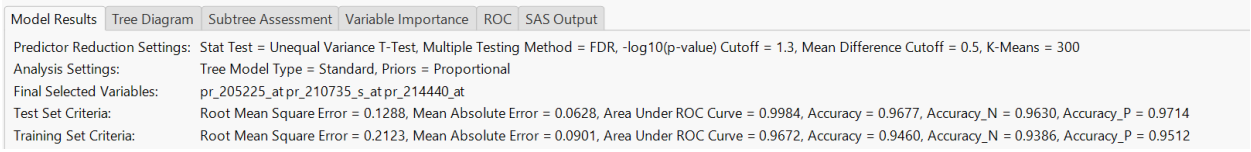


Analysis tab.

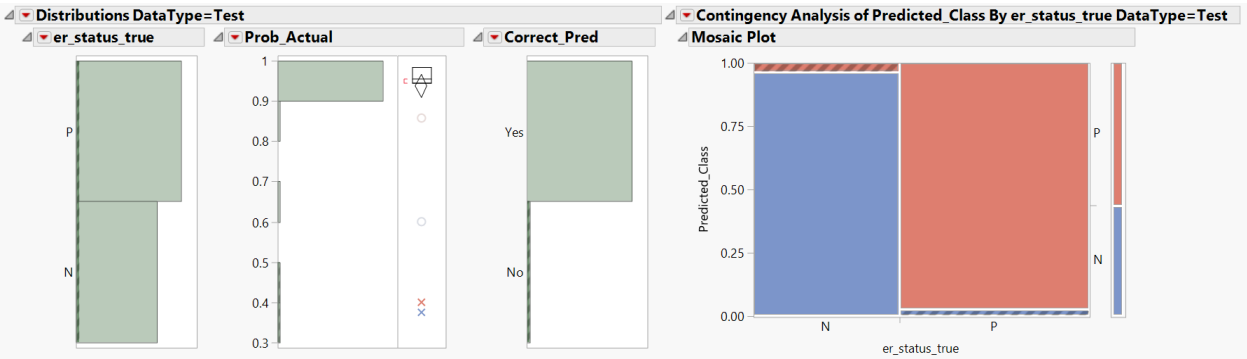
2. On the **Options** tab, select one of the **Test Data Sets** created in the *Subsetting a Test Data Set for Test Set Model Comparison*. We will use ***gse20194_wide_test_2.sas7bdat***.
    - An optional **Validation Data Set** can be added on the **Options** tab if desired.
3. Click **Run** to begin the analysis.

Results

4. When the analysis finishes running, the results will appear in a new window. At the very top of the **Model Results** tab are summary results for the model from the test set and training set. This section also contains the **Final Selected Variables** which shows us that there are 3 final probes included in the model.

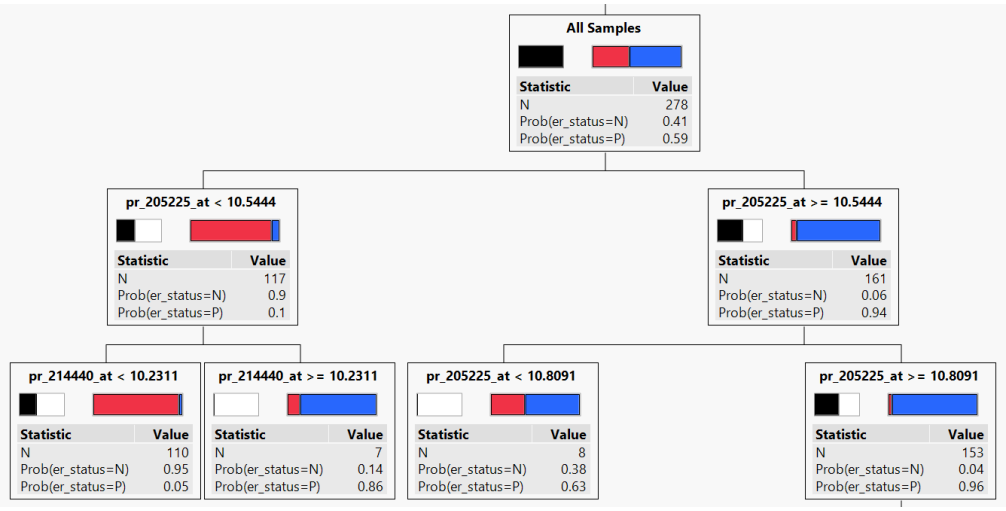


5. Beneath the summary of fit statistics there are distribution plots for the test set and then the training set. Beneath each set of plots are summary statistics.



- The first plot shows the distribution of the dependent variable. Recall that we designated estrogen receptor status as the dependent variable when creating the model. This shows the distribution of both positive (P) and negative (N) status.
- The middle plots both show statistics regarding the accuracy of the model. The **Prob_Actual** plot is the distribution of the posterior probability. Values below 0.5 here are a misclassification. The **Correct_Pred** is the simple distribution of correct vs. incorrect calls that the model made. You can see below the plot under the **Frequencies** tab, that the model was 96.7% correct for its predictions in the test set.
- The **Mosaic Plot** sorts all correct and incorrect predictions into cells for both classifications of er_status.

6. The **Tree Diagram** tab gives statistics for each of the three probes designated by the model including the number of data points for each level of each probe and the distribution of er_status for each level.

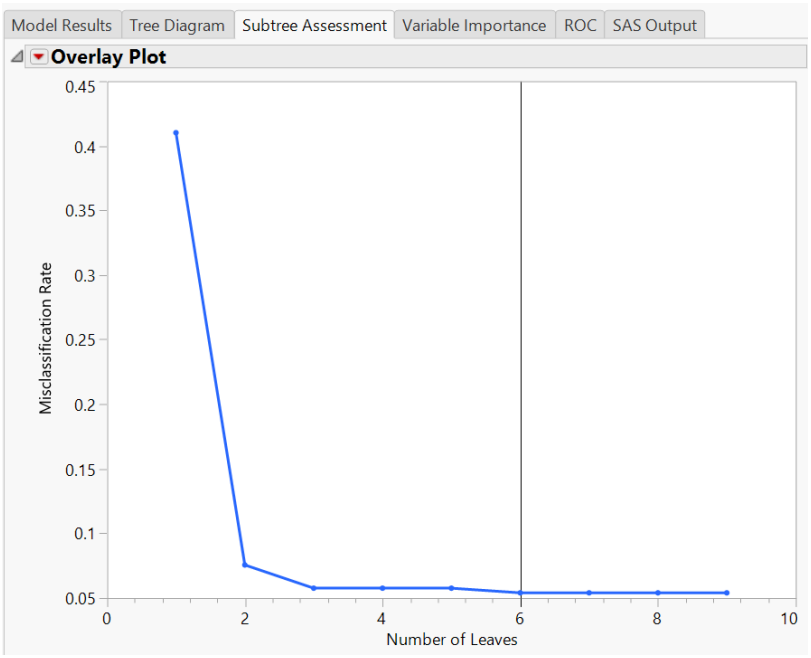


- The top of the **Tree Diagram** shows the probability of er_status for both N and P for the entire data set (278 samples). As the diagram moves further

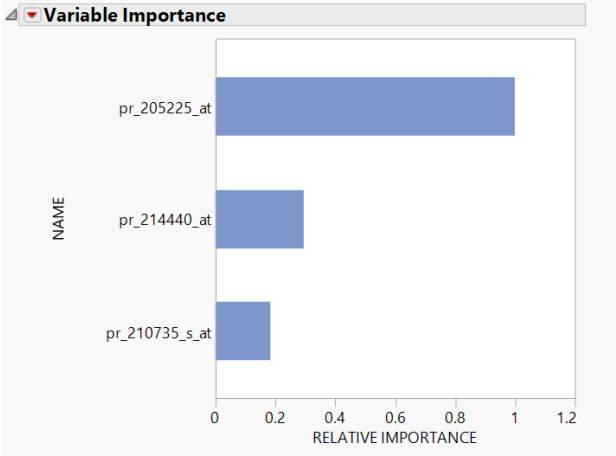
down, each branch is the statistically optimal way to segment the data for accurate prediction.

- The first level of branches has the probe “pr_205225_at”. This shows that the 117 samples with a value less than 10.54 for the probe have 0.9 probability of being negative for er_status, and for values greater than or equal to 10.54, samples have a 0.94 probability of being positive.
- Each subsequent branch of the tree further breaks up the samples from the branch above it.

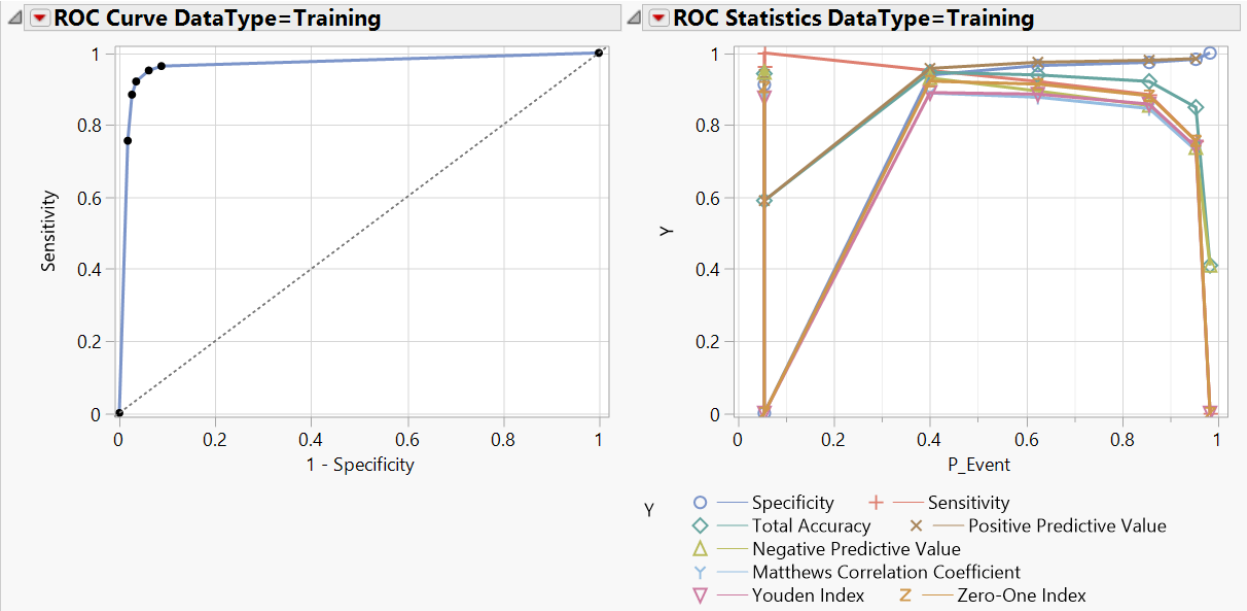
7. The **Subtree Assessment** tab plots the rate of misclassification vs the number of leaves in the tree. The vertical line is drawn at the actual number of fitted leaves in the model.
    - This model has 6 leaves, or segments of the tree that are not further segmented. The model classifies each sample into one of the 6 leaves before making a prediction.



8. The **Variable Importance** tab gives the relative importance for each of the three probes chosen as branches in the tree.
    - The probe “pr_205225_at” has the most influence on this model.



9. The **ROC** tab plots the plots the correctly identified proportion of true positives vs correctly identified true negatives for both the test and training set.
    - The area under the curve on the first plot is a measure of the sorting efficiency of this model. A value of 1 here would represent perfect sorting.
    - The second plot show several performance statistics for different posterior probability cutoffs on the x-axis.



Summary

This document served as a guide to fitting a **Partition Trees** model to a set of markers to predict estrogen receptor status in human breast cancer data. The model was chosen after comparison to another model through cross-validation (See *Cross Validation Model Comparison Step-Guide*). The **Partition Trees** model makes predictions by choosing significant markers and then segmenting the data based on

values for those markers. This specific model split the data into 6 groups before making a prediction, and was over 90% accurate.