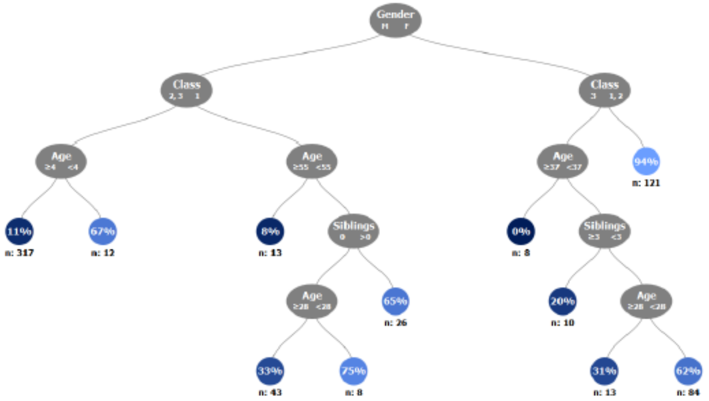
Decision Trees

- **Use:** Decision Trees are a non-parametric method used to predict a continuous or discrete output variable using continuous and/or binary input variables.

- **Fit:** Decision trees are created by recursively splitting the feature space to maximize information gain. The common splitting criteria decision trees are the Gini index or Entropy for classification, and the Mean Squared Error for regression. The most common implementation of decision trees is the CART algorithm.



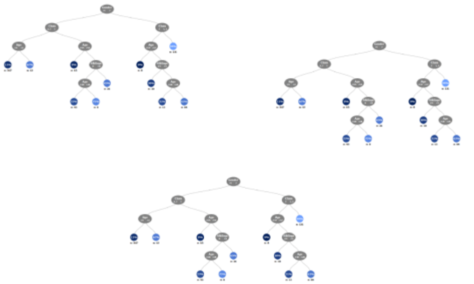
- **Input / Tuning Parameters:**
  - Max Tree Depth.
  - Minimum Samples in Leaf.
  - Minimum Information Gain.

- **Considerations**
  - The Decision Tree is a visually engaging model that can be easily interpreted.
  - The Decision Tree is a non-parametric technique that is relatively unaffected by outliers
  - The Decision Tree is a common base learner for powerful ensemble models such as Random Forests and Boosted Trees.
  - The Decision Tree will have high variance if it is not pruned.

Random Forests

- **Use:** Random Forests are an a non-parametric ensemble method used to predict a continuous or discrete output variable using continuous and/or binary input variables.

- **Fit:** A Random Forest is a collection of decision trees that have been created on different samples of the training data using random feature selection at each split. The underlying decision trees are not pruned.



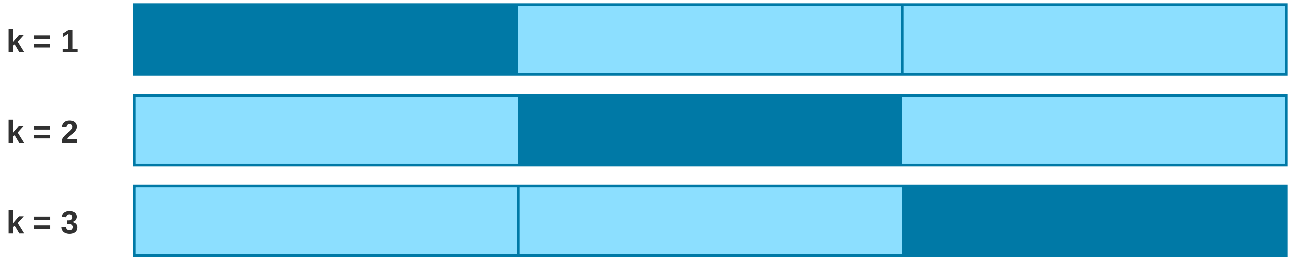
- **Input / Tuning Parameters:**
  - Number of trees
  - Number of candidate variables at each split

- **Considerations**
  - The Random Forest requires very little tuning to fit and thus often works well “out-of-the-box”.
  - Variable importance scores for the Random Forest are a popular method for variable selection
  - The Random Forests is not as interpretable as some of the other methods such as Decision Trees and Linear Regression.

Parameter “Optimization”

- **Use:** Model tuning is the process of evaluating the performance of a model across many values of its input parameters to find the “optimal” values.

- **Cross-Validation:** A popular technique that partitions data into sections to evaluate the ability of a model to generalize to out of sample observations. See example of 3-fold validation below.



- **Metrics for Classification:**
  - Accuracy
  - Receiver Operating Characteristic Area Under the Curve (ROC AUC)

- **Metrics for Regression:**
  - Root Mean Squared Error (RMSE)
  - R^2