To perform classification, we will be implementing decision trees and bagging. Each of the decision trees will use the genotypes of the 1007 individuals for the 3672 SNPs from four different genes to generate branches. To reduce computation time both the SNPs with very rare (<0.001) and very common (>0.999) genotype frequencies were removed as they likely won’t be of a significant benefit during classification. Additionally, a portion of the data will be set aside as training data, and the remainder of the observations will be used during the prediction stage. Bagging is being used to increase accuracy and ensure the overall model will be a suitable classifier on new data. For this analysis, 300 different trees will be generated, and the average predictive performance will be represented by a receiver operating curve (ROC) and the resulting area under the curve (AUC).

Diagram

Description automatically generated Bagging generated 300 trees, an example of which are visualized in figure 1. Additionally, a singular tree was generated for comparison purposes. Interestingly, many of these trees had 3 or more SNPs when a tree generated without bagging only used 1 SNP for classification.

After rounding the average bagging predictions to fit within the binary classification, bagging resulted in 4 misclassifications. To visualize prediction abilities an ROC plot was used, as seen in figure 2. Bagging has a great trade-off between sensitivity and specificity with an AUC value of 0.9998.

Figure : Examples of trees generated during the bagging procedure

Chart

Description automatically generated

Overall, given the nature of the data a singular decision tree would be an acceptable way to classify this data. Although, the bagging trees provided a higher AUC value and therefore are considered a better classifier. Bagging is also a good practice to implement as it improves accuracy and reduces the chances of overfitting to a given set of data.

Figure 2: ROC plot for bagging trees which shows the trade-off between sensitivity and specificity for the final predictions