White Wine Classification

J Steven Raquel
PSTAT 131
Spring 2017

Questions to Answer

- What is the best method of classification between decision tree, k-nearest neighbors, and randomForest?
- What are the most important predictors when it comes to estimating whether a wine is “good”?

The Data

- “Wine quality” dataset from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Wine+Quality)
- 12 attributes:
  - 11 numeric - fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol
  - 1 categorical - quality
- 4898 observations
- Some numeric predictors are on a different scale (e.g. pH and density)

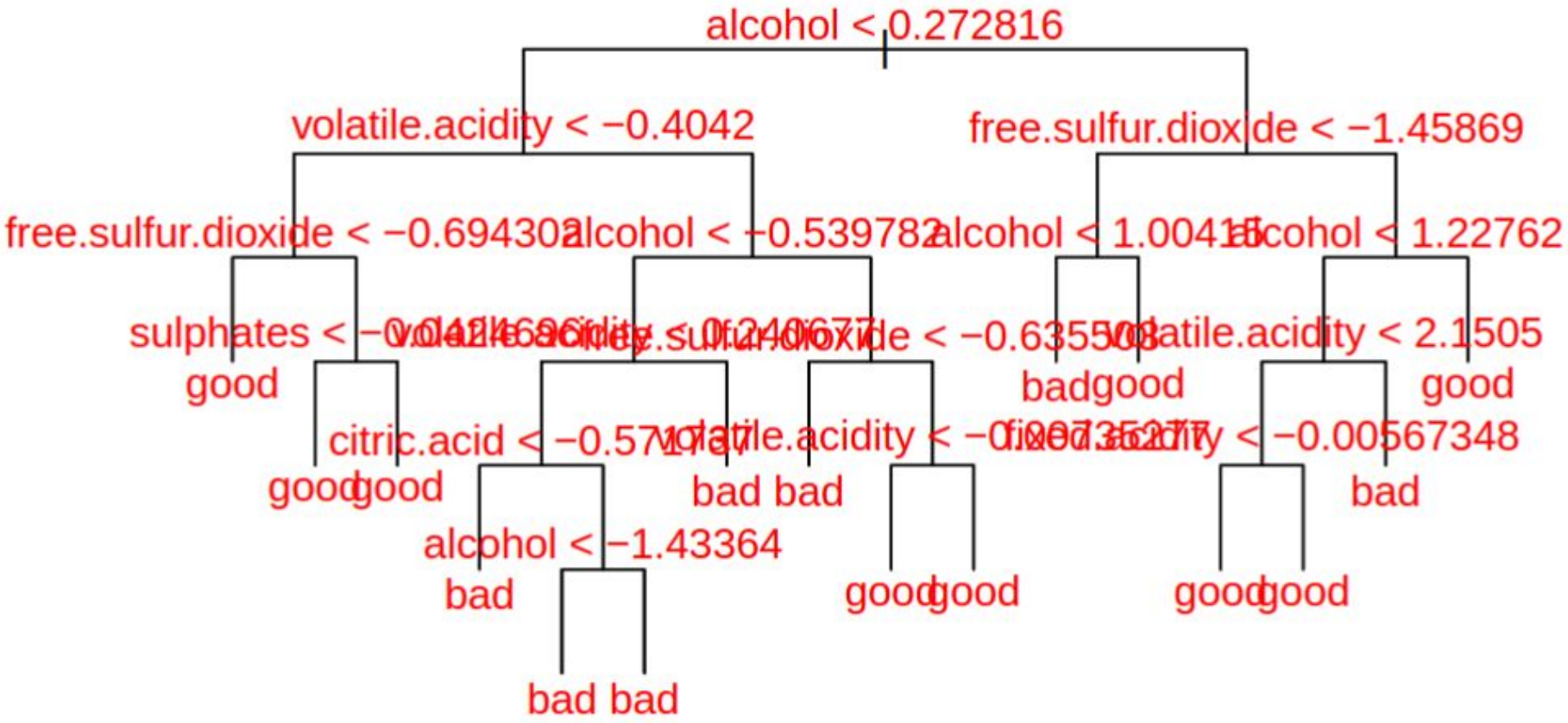
Pre-Processing

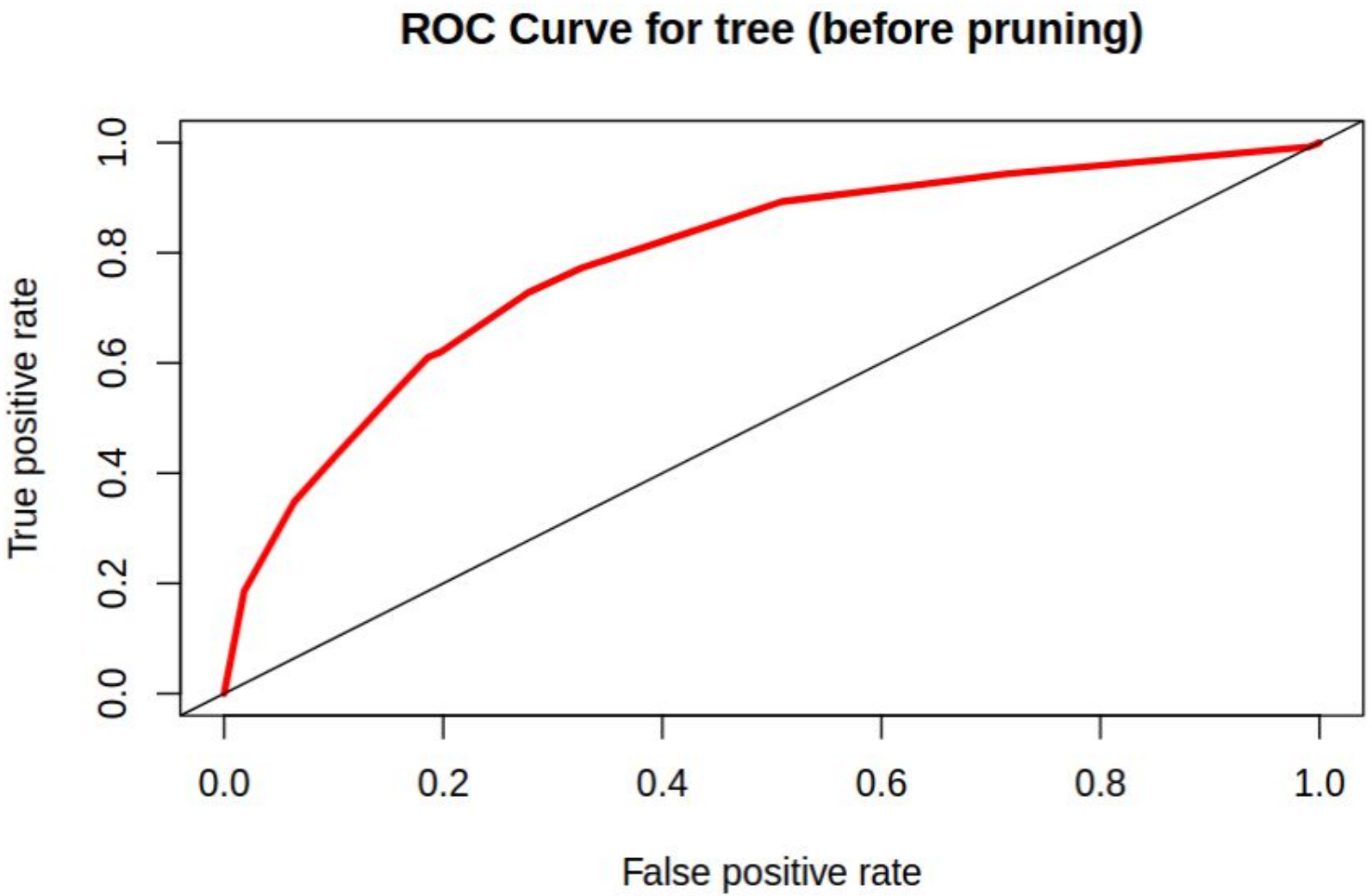
- Density was strongly correlated with residual sugar and alcohol, and moderately correlated with total sulfur dioxide.
- Total sulfur dioxide and free sulfur dioxide were also moderately correlated.
- We decided to **drop** residual sugar, density, and total sulfur dioxide to address *multicollinearity.*
- We **scaled** the remaining **8** numeric predictors, and converted the quality variable into a two-level categorical variable titled '**label**' with two levels: **good** and **bad**.
- Split the dataset into a 1000 observation test set, and a 3898 observation training set.

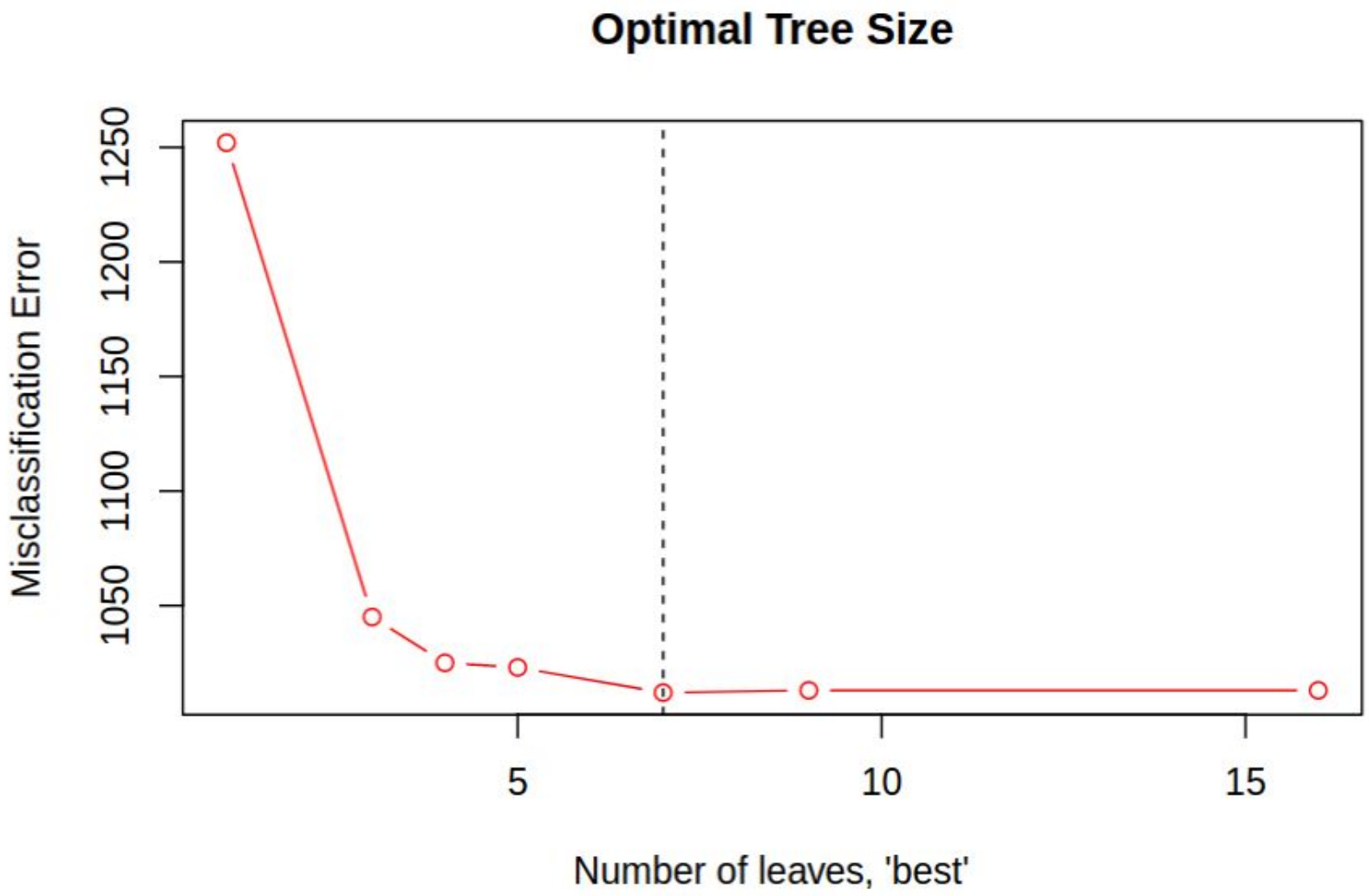
Decision Tree

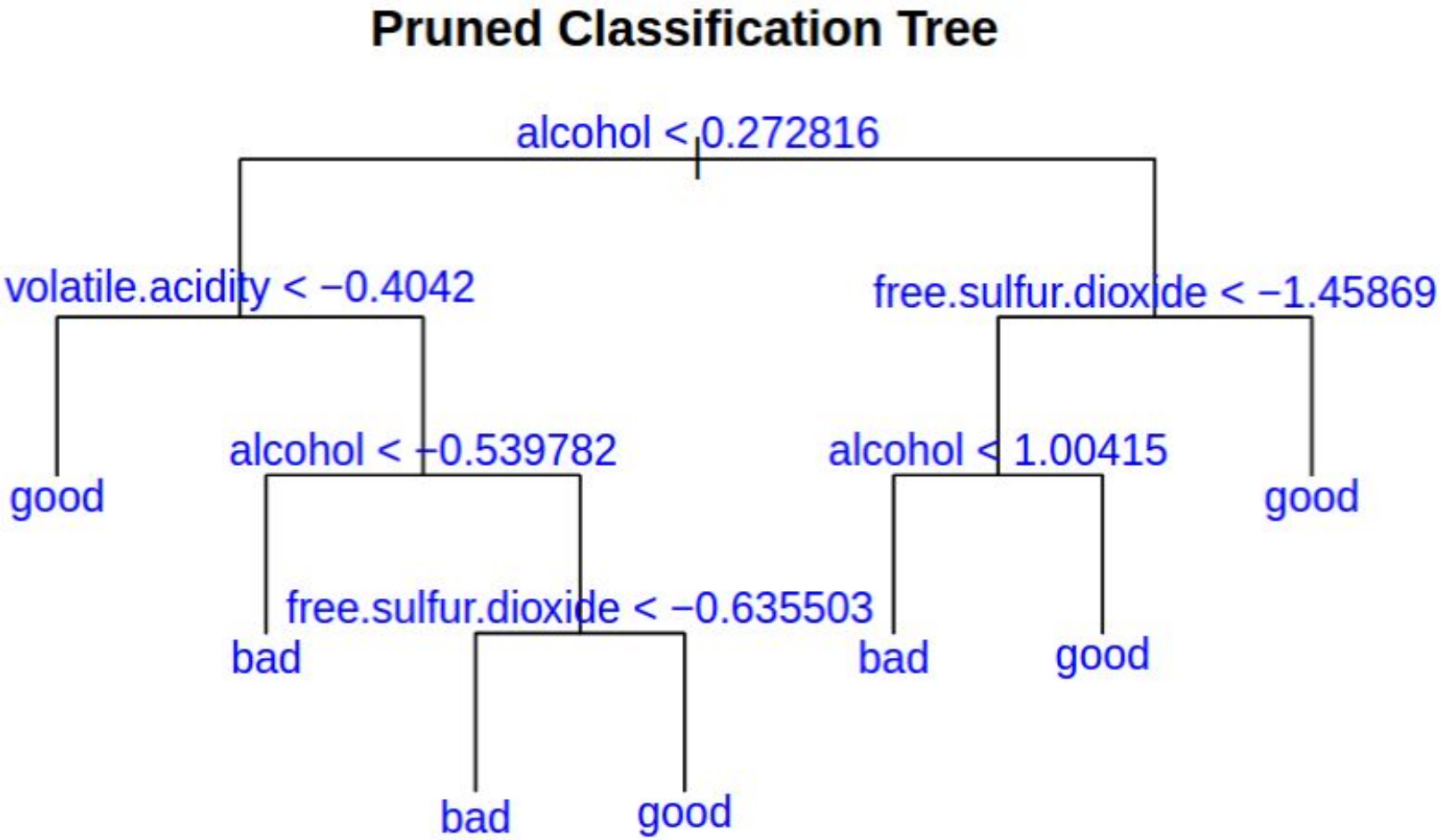
- The tree model with the formula with all 8 predictors, only utilized **6**: alcohol, volatile acidity, free sulfur dioxide, sulphates, citric acid, and fixed acidity.
- 16 terminal nodes
- Metrics:
  - Accuracy rate: **0.748**, Error Rate: **0.252**, AUC: **0.788**
- We used **10-fold** cross validation and determined that the best size of the tree is **7 nodes**.
- After pruning the tree, the accuracy rate did not change, but the AUC decreased.
  - Accuracy rate: **0.748**, Error Rate: **0.252**, AUC: **0.756**

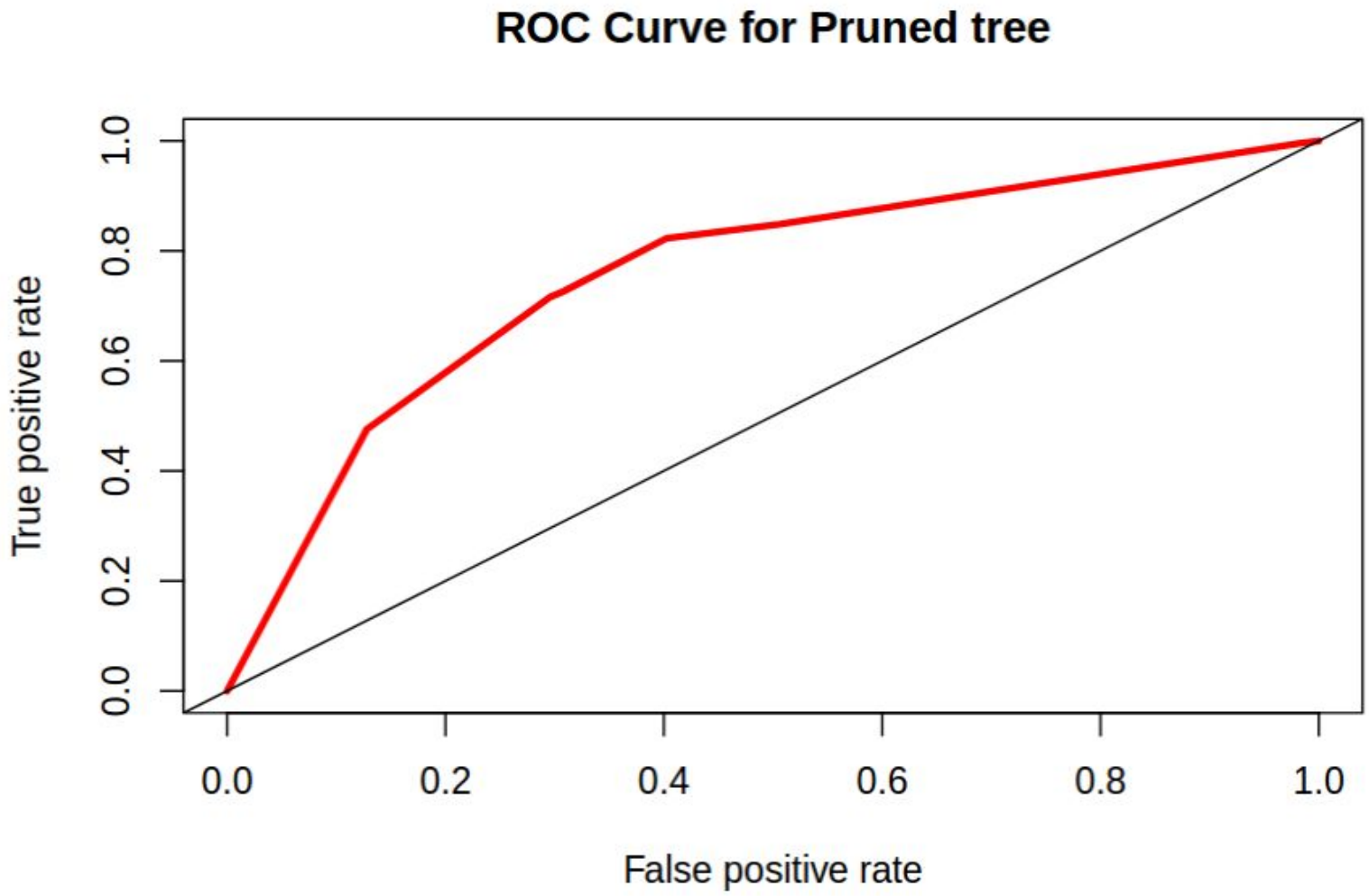
Classification Tree (Before Pruning)




ROC Curve for tree (before pruning)
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

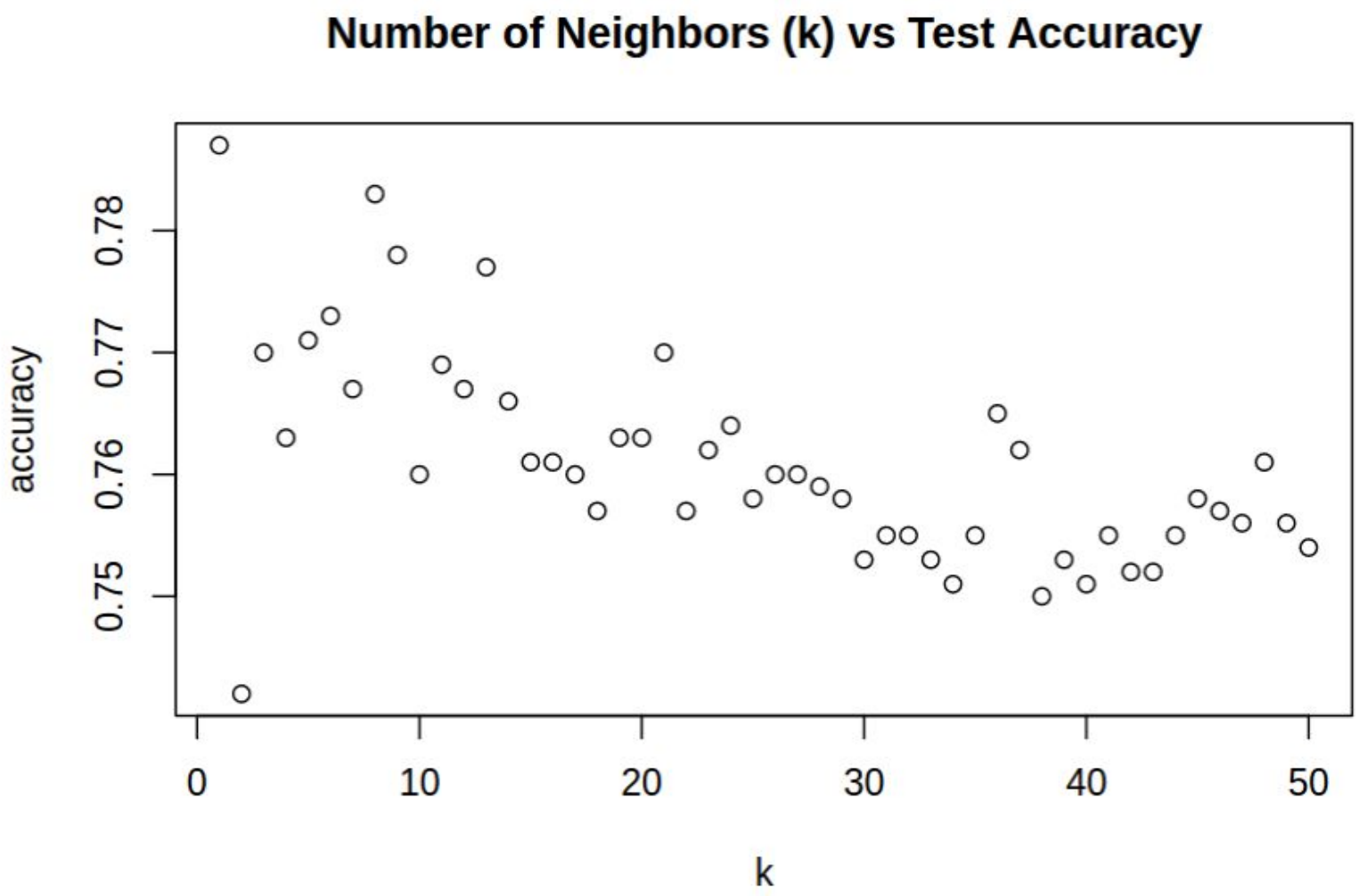

Optimal Tree Size
Misclassification Error
1050
1100
1150
1200
1250
5
10
15
Number of leaves, 'best'

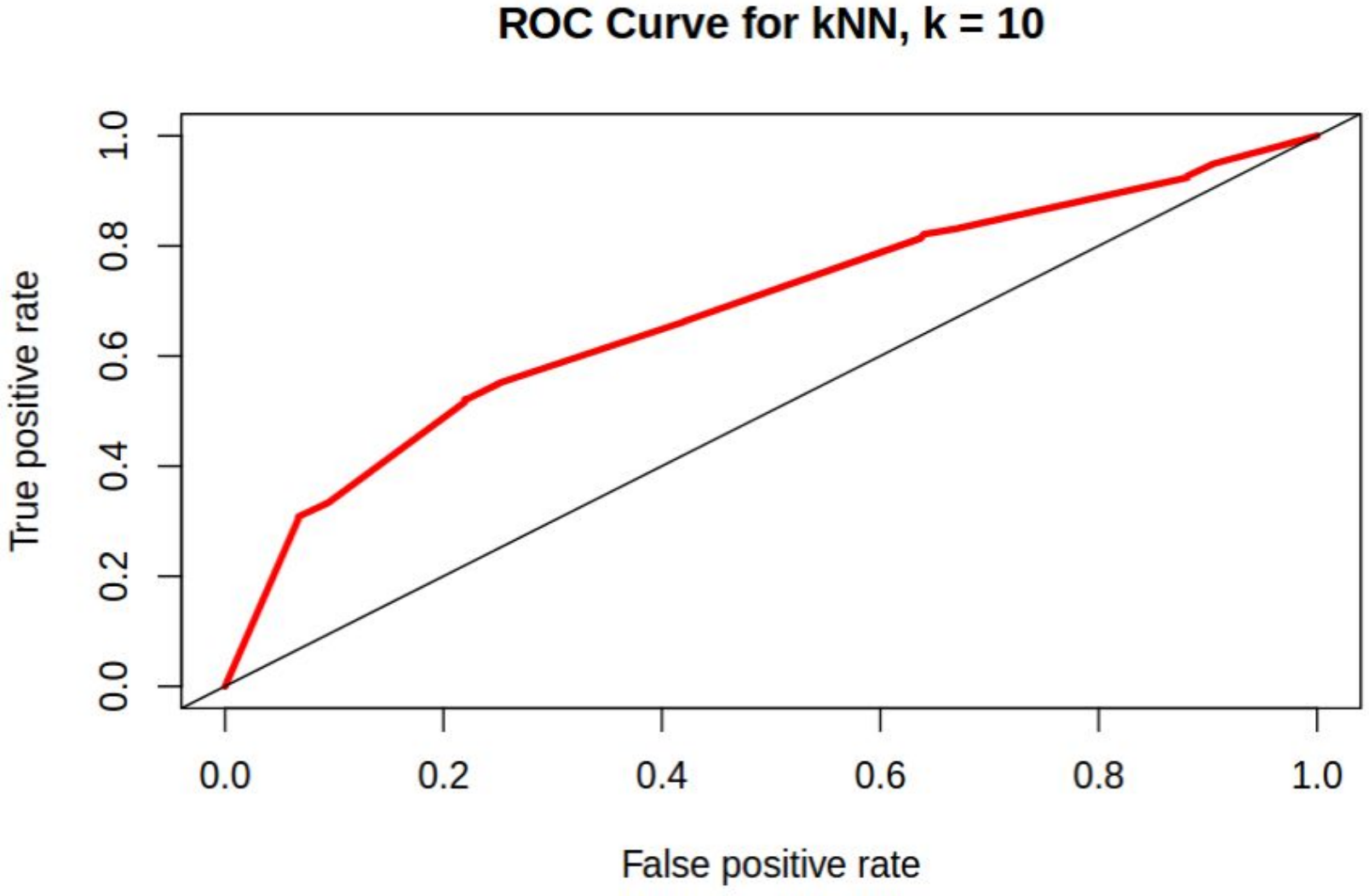

Pruned Classification Tree
alcohol < 0.272816
volatile.acidity < −0.4042
free.sulfur.dioxide < −1.45869
good
alcohol < −0.539782
alcohol < 1.00415
good
free.sulfur.dioxide < −0.635503
bad
bad
good
bad
good

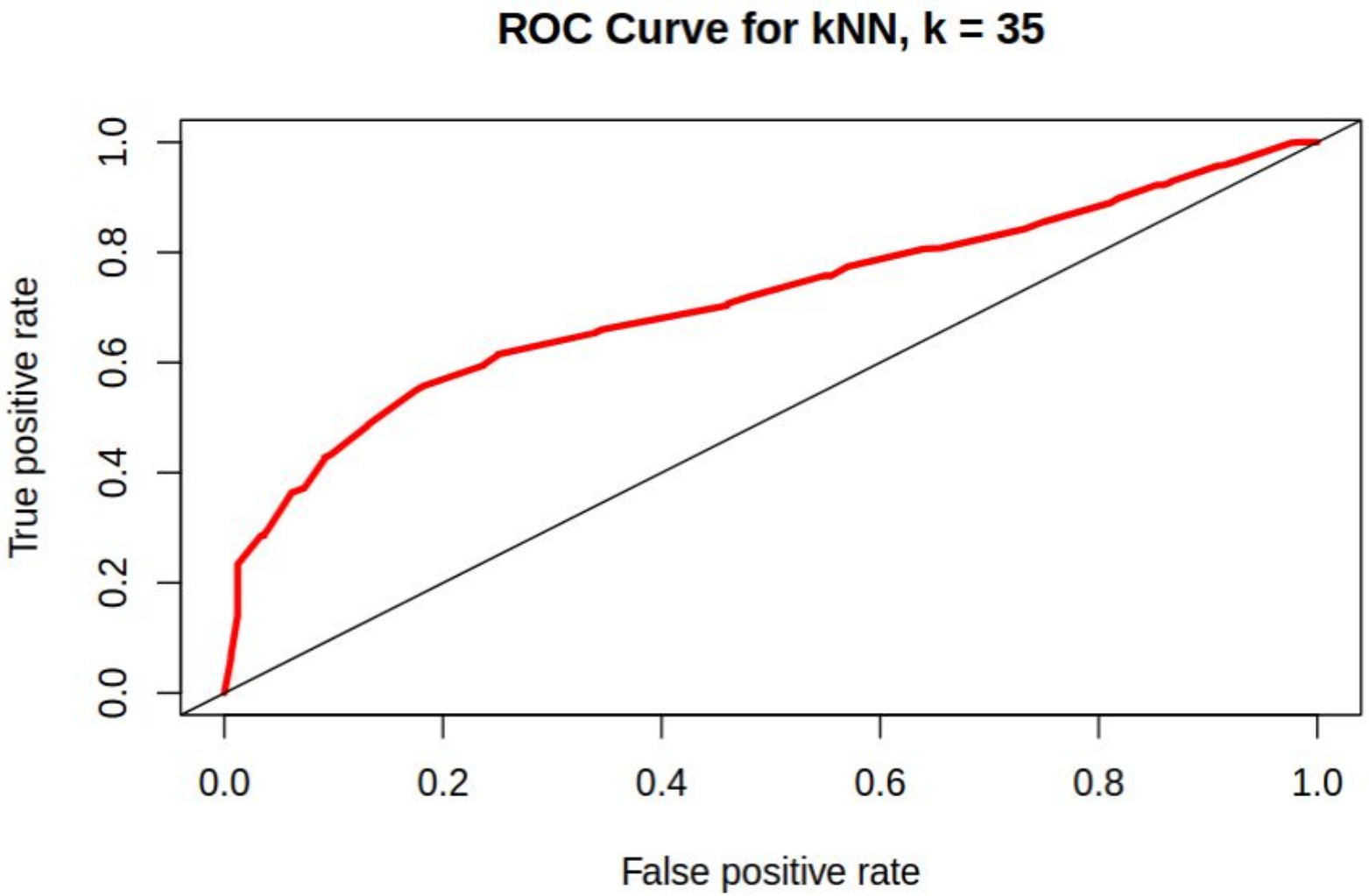

ROC Curve for Pruned tree
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

k-Nearest Neighbors

- Started with k = **10** nearest neighbors
  - Accuracy: **0.765**, Error rate: **0.235**, AUC: **0.679**
- We found that accuracy was highest with k = 1, but this has very high variance and is prone to overfitting.
- We opted for k = **35** to still get a good accuracy but without so high a variance.
  - Accuracy: **0.755**, Error rate: **0.245**, AUC: **0.711**

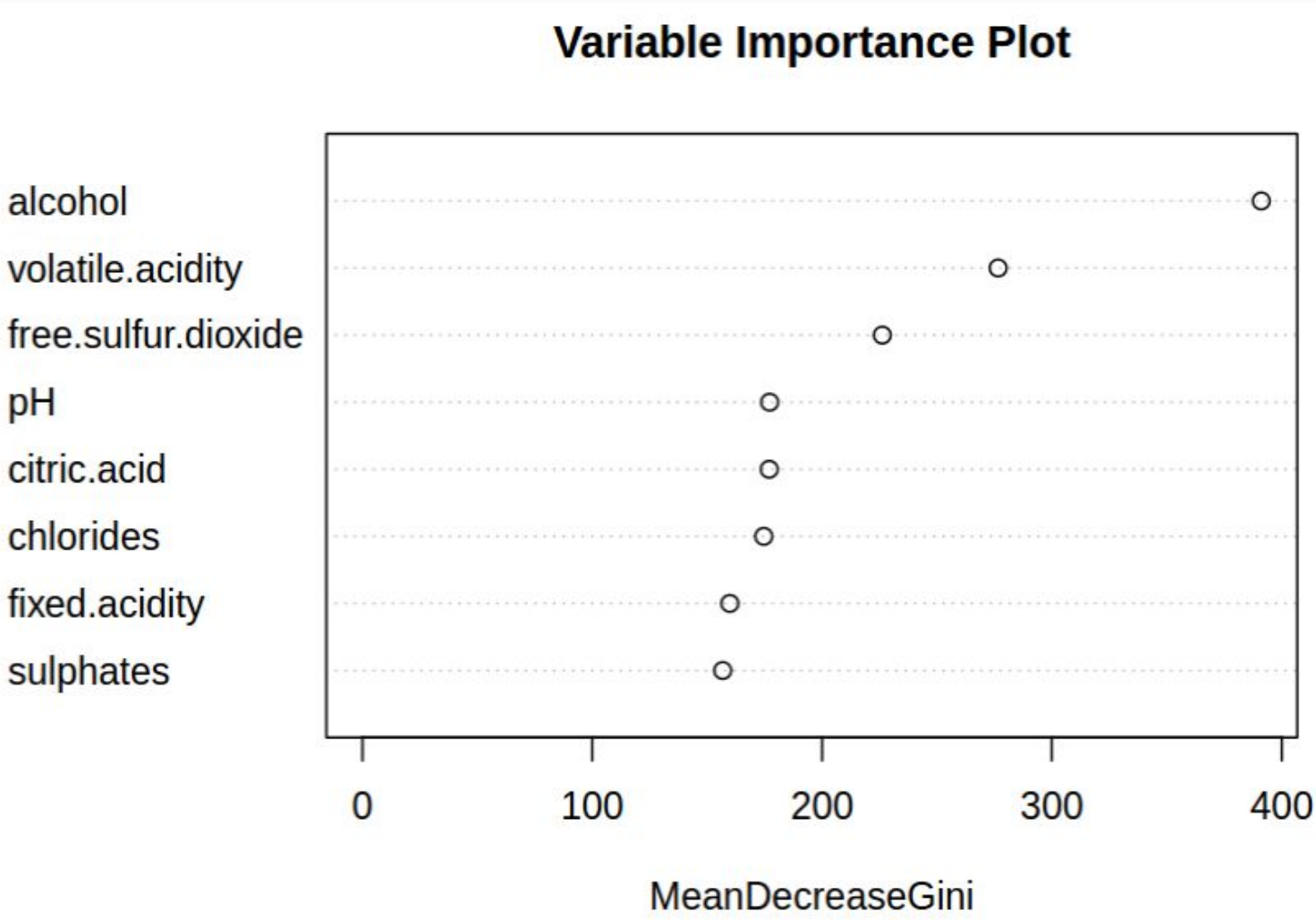

Number of Neighbors (k) vs Test Accuracy
accuracy
0.75
0.76
0.77
0.78
0
10
20
30
40
50
k

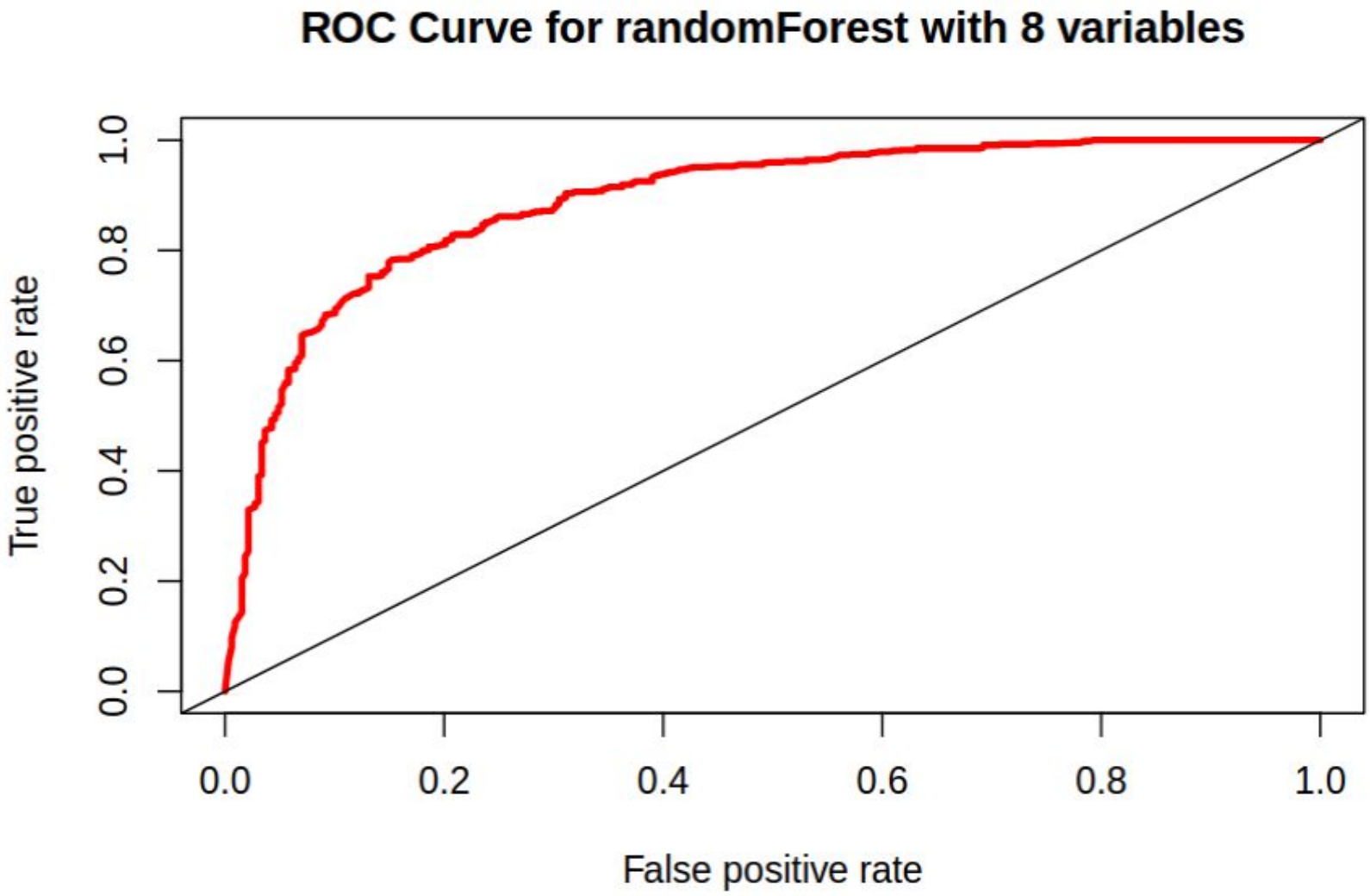

ROC Curve for kNN, k = 10
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

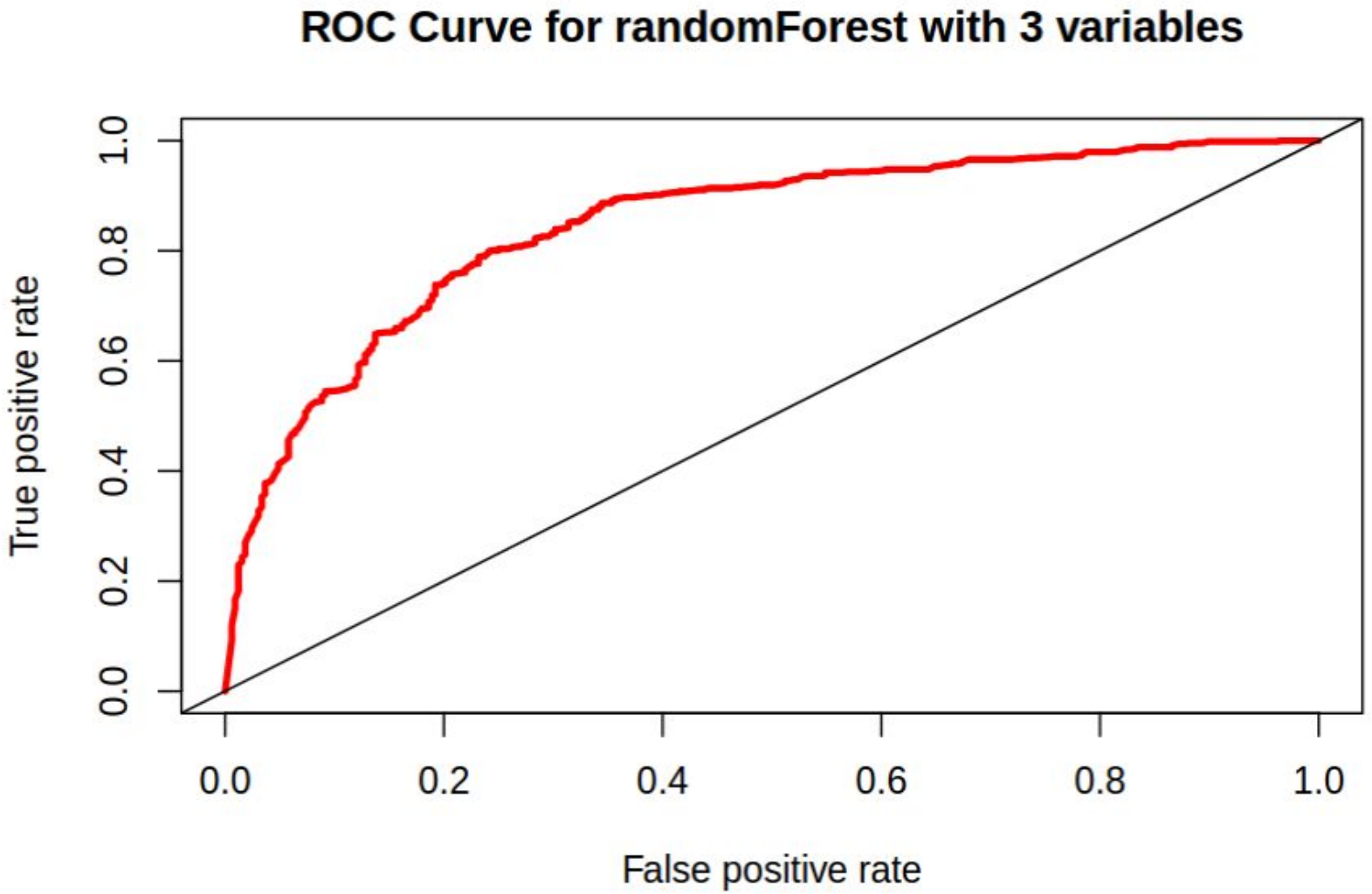

ROC Curve for kNN, k = 35
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

randomForest

- Uses bootstrap aggregating (bagging) to create a ‘forest’ of decision trees
- Chooses the classification that is most commonly chosen by the many decision trees.
- Introduces randomness that minimizes overfitting, unlike one decision tree.
- We first used all 8 predictors:
  - Accuracy: **0.817**, Error rate: **0.183**, AUC: **0.889**
- Later used only the 3 most important predictors: alcohol, free sulfur dioxide, volatile acidity
  - Accuracy: **0.806**, Error rate: **0.194**, AUC: **0.846**


Variable Importance Plot
alcohol
volatile.acidity
free.sulfur.dioxide
pH
citric.acid
chlorides
fixed.acidity
sulphates
0
100
200
300
400
MeanDecreaseGini


ROC Curve for randomForest with 8 variables
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve for randomForest with 3 variables
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Comparison of Models

Model	Accuracy Rate	Error Rate	AUC
tree	0.748	0.252	0.788
pruned tree	0.748	0.252	0.756
k=10 kNN	0.765	0.235	0.679
k=35 kNN	0.755	0.245	0.710
full randomForest	0.817	0.183	0.889
small randomForest	0.806	0.194	0.846

Conclusions

- Of the three classification methods (decision tree, k-nearest neighbor, and randomForest), **randomForest** was the most accurate.
- Aside from being the most accurate, randomForest does not face the issue of overfitting that decision trees tend to run into, and kNN has a weaker AUC than both.
- Pruning the decision tree gave us the idea that the three most important variables are **alcohol**, **free sulfur dioxide**, and **volatile acidity**, which was corroborated by the randomForest model.
- The less complex model randomForest model that utilized these three variables was slightly weaker than its more complex counterpart but still better than decision trees and kNN.

Outstanding Questions

- Which of these methods are most prone to over or underfitting to the training data?
- What is the ideal amount of alcohol to put into a wine to classify it as “good”?
- What would be the best k in k-nearest neighbors to minimize variance AND bias?
    - Same question, but for terminal nodes in a decision tree
- Is dropping variables in randomForest necessary?