HarvardX Data Science Capstone: Classifying Wine

Michael Lewis

19 September 2019

Executive Summary

The goal of this project was to evaluate and classify the wine of three cultivators from the same region in Italy. Each wine had thirteen physical or chemical properties measured. These were used to train machine learning algorithms to distinguish between each cultivator, or "type". The data used for this project came from the University of California at Irvine's Machine Learning Repository.

In order to determine a rigorous analytical approach, exploratory data analysis was performed and the data integrity was evaluated. From there, a validation set was generated via holdout sampling, and data transformations were made. Next, several machine learning algorithms (CDT, Random Forest, and KNN) were trained and dimension reduction (PCA) explored. Each algorithm's perfromance was compared, and though all performed well, the decision was made to move-forward with the K-Nearest Neighbor algorithm. When applied to the validation data, this model acheived 94% accuracy while maintaining a relatively intuitive understanding of how the algorithm determines the classification of each wine. The thirteen attributes measured and the described approach constitute a success for classifying this region's wine types. However, further work (with a larger sample of wines from diverse regions) is needed to enable development of a generalized wine classification algorithm.

Methods & Analysis

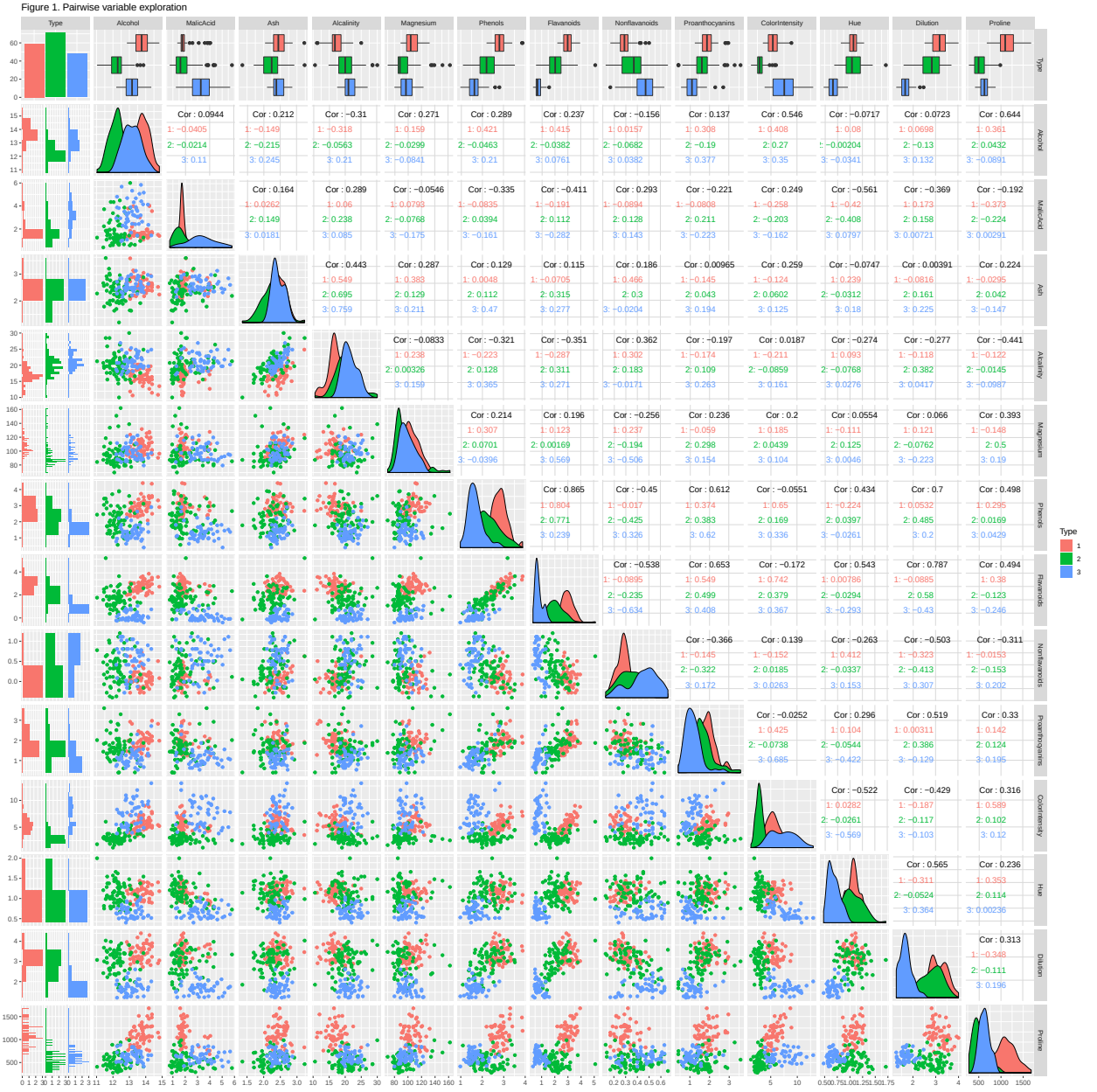
In order to become more familiar with the data set, and glean insight as to how to proceed with a classification algorithm, exploratory data analysis was performed. First, the classification variable ("type") was updated to a factor and missing values were searched for - none were found. Next, coupling variable summary statistics and their pairwise comparisons showed variable standardization would likely benefit prediction efforts as some algorithms use Euclidean Distance which is affected by "scales" of variables. Other interesting insights revealed by evaluating individual plots of the pairs-plot included:

(1) Among some of the predictor variables, there is some substantial correlation [e.g. Flavanoids-Proanthocyanins (0.653) and Flavanoids-Phenols (0.865)]. Thus dimension reduction might prove fruitful. These correlations are respectivley shown in **Figure 1** at the following indeces: [10,8] and [8,7].

(2) High accuracy is a possibility given the ability to delineate based on the relatively stark stratifcation of wine Type 1 across Proline as well as wine Type 3 along Malic Acid and Color Intensity. These distributions are respectivley shown in **Figure 1** at the following indeces: [14,14], [3,3], and [11,11].

Next, to mitigate the risk of overfitting and allow for the evaluation of model performance, the data set was partitioned into training (60%), test (20%), and validation (20%) sets. From here, the predictors were standardized by subtracting the mean of a given variable from each observation, then dividing the result by the variable's standard deviation.

Figure 1. Pairwise variable exploration



Next, three algorithms and four models were trained then evaluated on the test data. In each case, model performance was determined by examining accuracy. This measure is reasonable as the cost associated with false negatives and false positives is equivalent.

First, a Classification Decision Tree was trained with a tuned complexity parameter. **Figure 2** shows the modest effect tuning the complexity parameter has on accuracy. **Figure 3** shows the final model.

Figure 2. Complexity Parameter Tuned for Accuracy

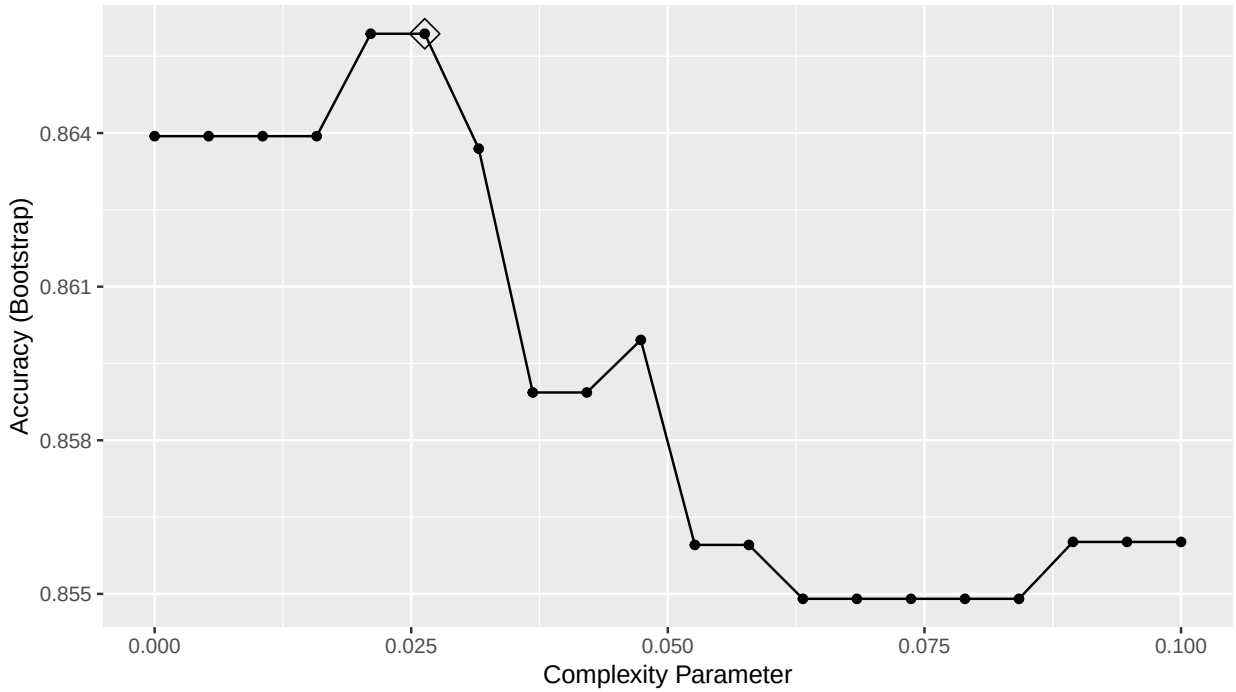
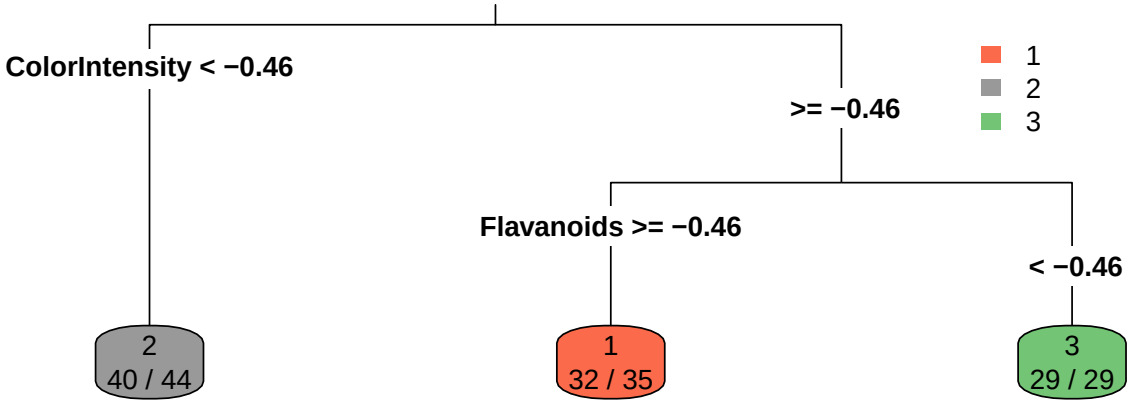
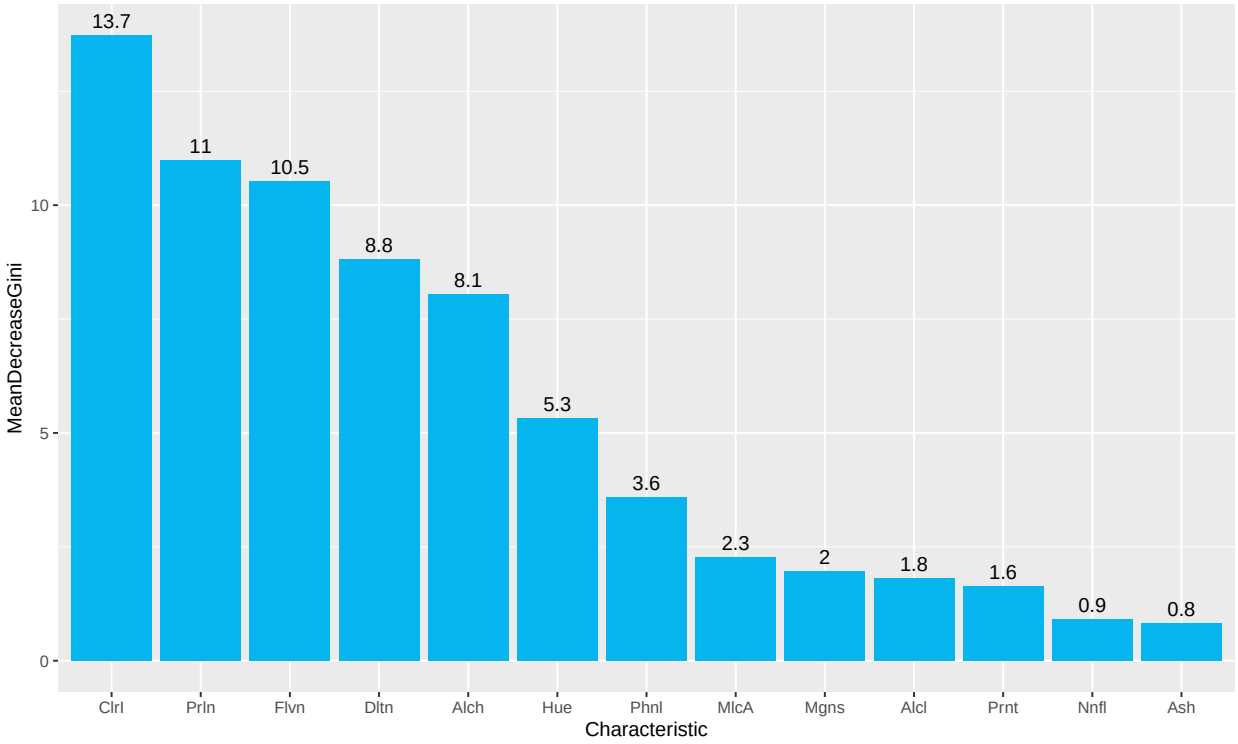


Figure 3. Wine Type Classification Tree



Second, a random forest model was trained and the variable importance extracted. **Figure 4** shows the relatively large roles that color intensity, proline, flavanoids, dilution, and alcohol content played in decreasing node impurity.

Figure 4. Variable Importance from Random Forests



Lastly, two K-Nearest Neighbor (KNN) models were trained. The first was trained using the original 13 variables. **Figure 5** shows the effects of the number of neighbors on accuracy. Given the relatively distinct nature of each wine 'type' and the correlation of several explanatory variables, dimension reduction was performed before training the second KNN model. Principal Component Analysis (PCA) was performed and the number of components selected for the model was based on the second elbow in a scree plot. **Figure 6** shows the scree plot of the components. Similarly, **Figure 7** shows the and the cumulative variance explained by the components. The first 6 components explain 86.3% of the variance contained in the original 13 variables. **Figure 8** shows the optimal tuned number of neighbors for the KNN-PCA model.

Figure 5. Neighbors Tuned for Accuracy

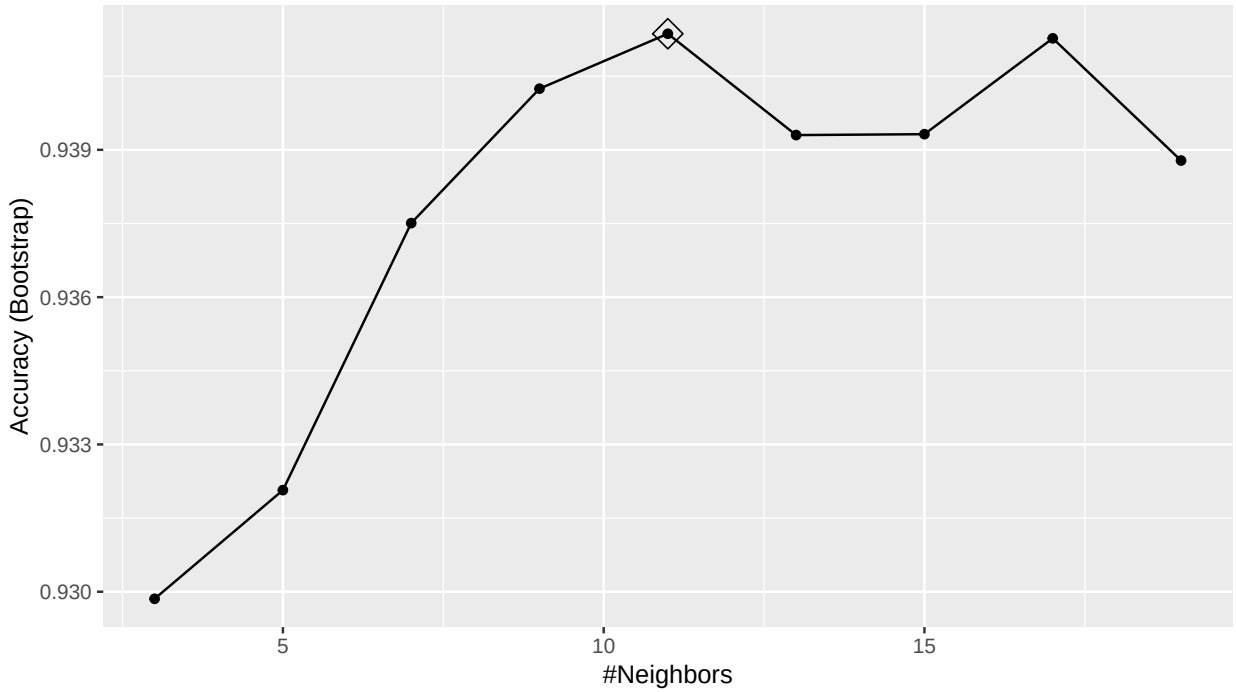


Figure 6. Scree plot − Wine PCA

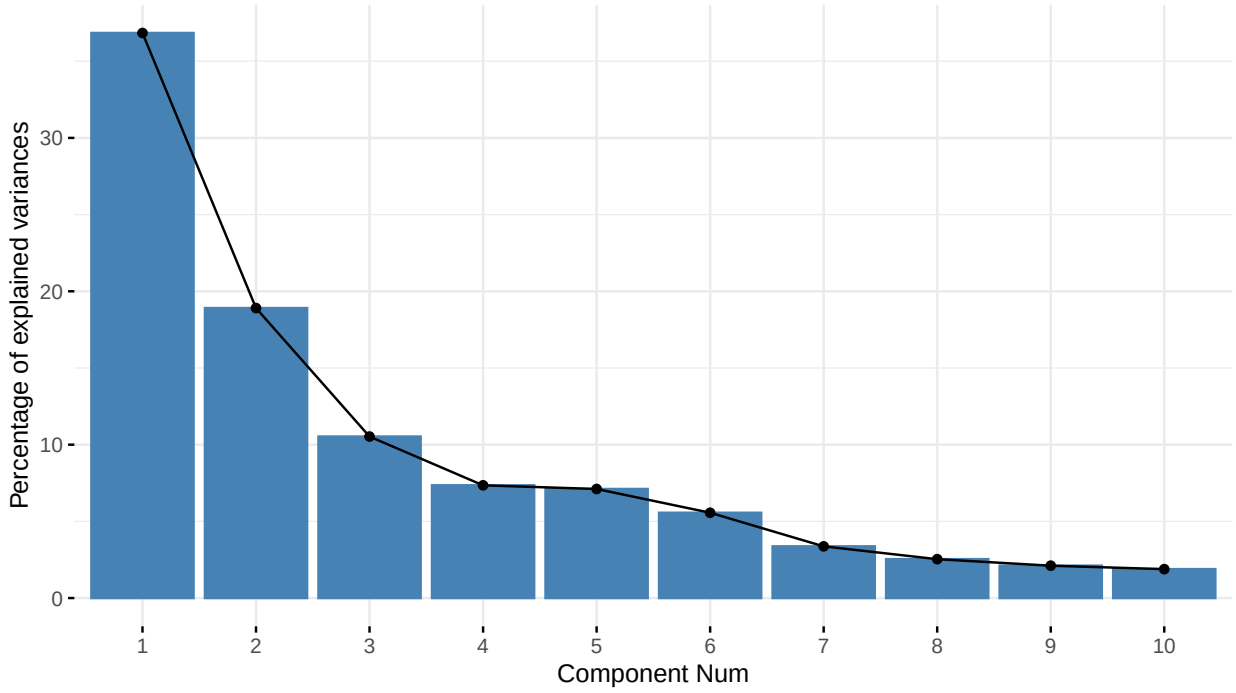


Figure 7. Cumulative Variance Explained in PCA

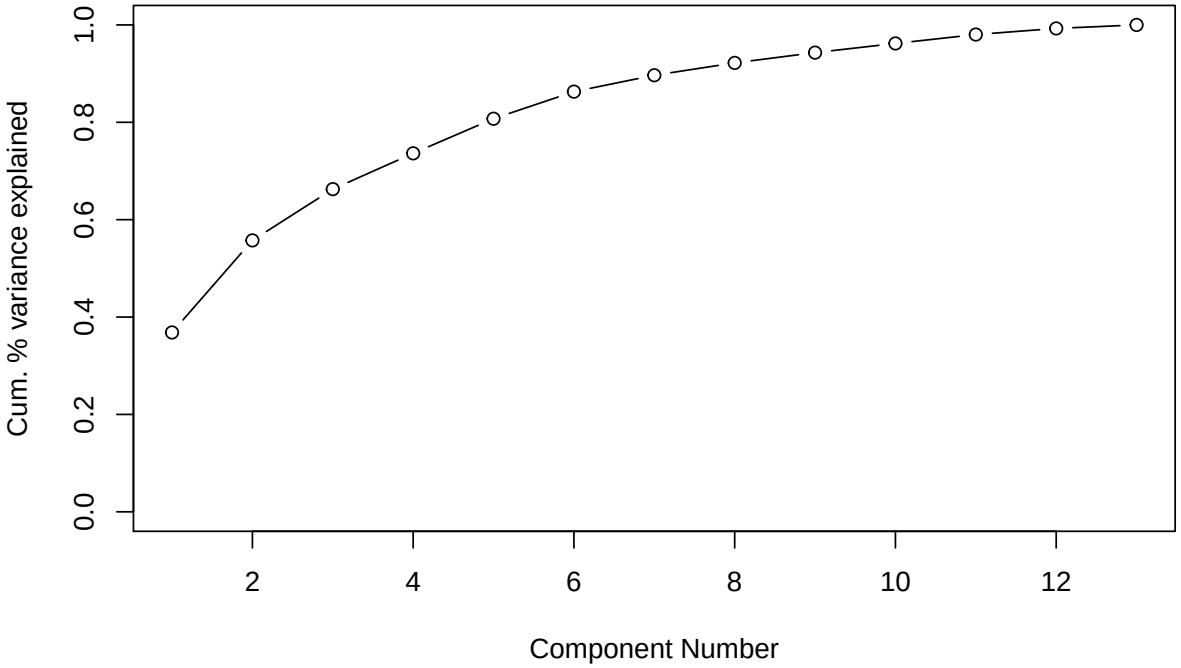
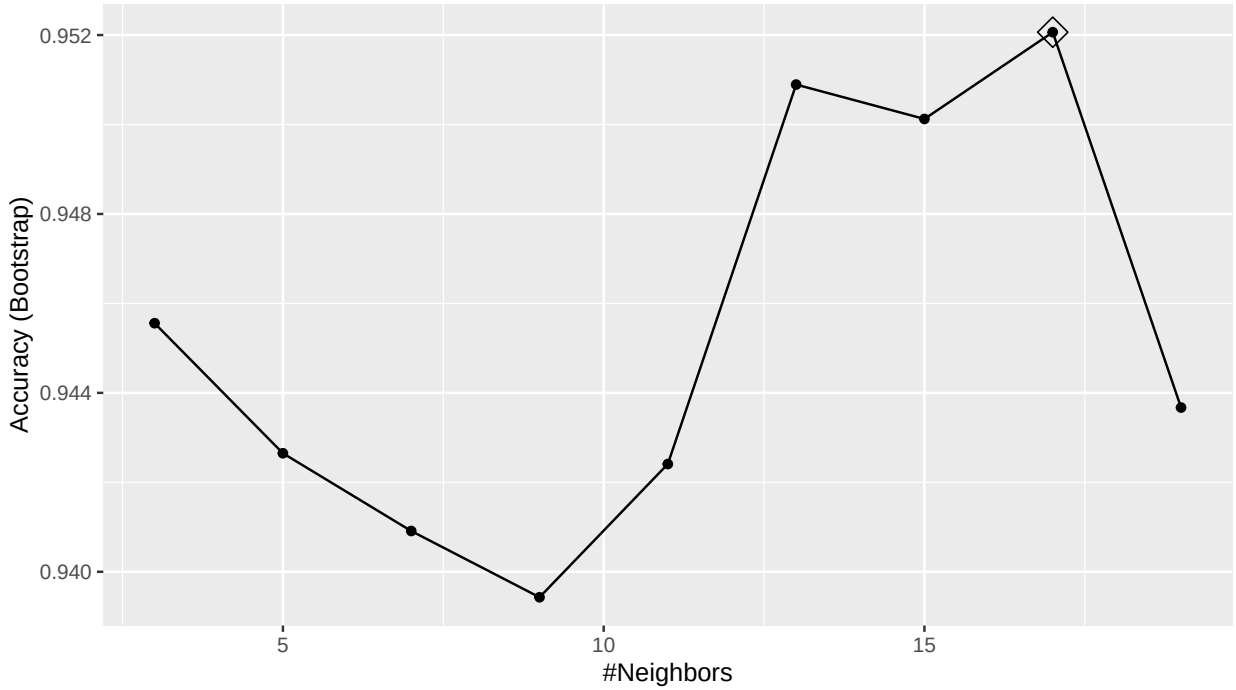


Figure 8. Neighbors Tuned for Accuracy under PCA



Results

Table 1 shows the accuracy of each model when evaluated against the test data. As suggested during the exploratory data analysis phase, the stratification of wine types across the variables measured results in accuracy levels traditionally associated with model over-fitting (e.g. the lowest accuracy among all models was 94.4%).

Table 1: Comparing Model Accuracies

Methods	Accuracy
Decision Tree	94.4
Random Forest	100.0
KNN	97.2
PCA-KNN	100.0

This situation presents a unique, albeit somewhat uncommon challenge: What should the decision rule be when deciding between comparably high-performing models? In order to maintain straight-forward interpretability and avoid models with the highest likelihood of overfitting, **the standard K-Nearest Neighbor model** was selected for use on the "new" (validation) data. This model produces a final **accuracy of 94.1%**.

Conclusion

The challenge presented with this data set was to classify Italian wines by cultivator ("type"). The chemical and physical properties measured for each wine provided sufficient between-type variance to allow for the development of highly accurate models across three algorithms - Classification Decision Trees, Random Forest, and K-Nearest Neighbors. High accuracy was observed for models trained on the original 13 variables as well as the one trained on the first 6 principal components. Given its performance and parsimony, the unreduced K-Nearest Neighbor algorithm served as the ideal candidate to be evaluated against the holdout sample. It produced an accuracy of 94.1%.

Two important limitations of this project are: (1) The fact that all wines were from the same region in Italy, and (2) fewer than 200 wines were evaluated. In order to produce a more generalizable model, more regions and wines should be included. This is of particular interest for future projects that could be developed from such modeling - i.e. classifying fradulent wines. While the approach would be different than the one employed here, having a robust model of regional wine types would provide a baseline from which fraud detection could begin.