Experimenting with the performance of various feature selection methods on a wine dataset

Kevin Chen
University of California, Irvine
chenkj3@uci.edu

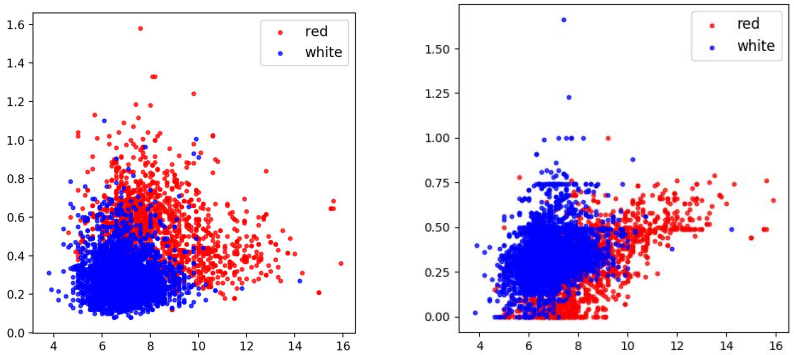
Zhenxiao Lin
University of California, Irvine
zhenxil1@uci.edu

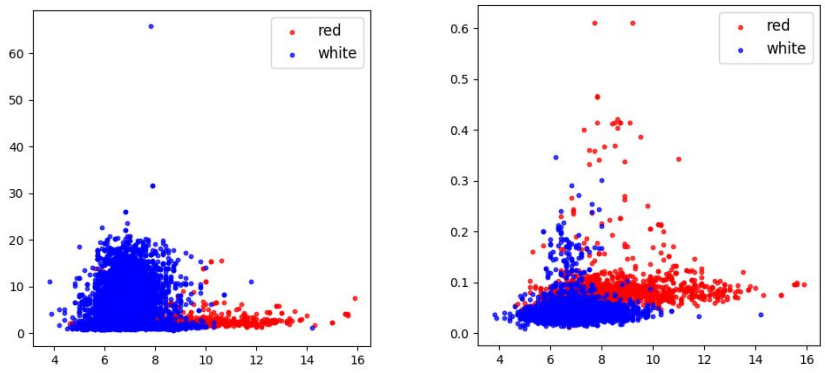
Daniel Kim
University of California, Irvine
kimd13@uci.edu

Data: Exploration and Preprocessing

The dataset that we chose to work with was the Wine dataset. The wine dataset has 11 features. 10 of these features measured the chemical makeup of the wine including the acidity, sugar level, and alcohol content. The last feature was an integer number that represented the quality of the wine on a scale of 1 to 10. We wanted to for classification of red or white wine using these features. In order to do this, we added another feature called “classification” that represented red or white wine. Red wine was represented with a value of 0 and white wine was represented with a 1.

To further explore the dataset, we decided we wanted a visual representation to aid in eliminating any data points or features. Since we were classifying red and white wine, we decided to graph features that compared these two types of wine. Scatter plots were created from taking the first feature and graphing it against the rest of the features. Below is a picture of the first four graphs.





The rest of the graphs depicted a conclusion that was very similar to those above. The blue data points (white wine) often hugged the bottom left corner more tightly while the red data points (red wine) always bordered the right side of the blue cluster and was always more spread out.

Based on seeing this situation for all features, we decided that we should not rule out any of the features, at the beginning. Instead, we wanted to perform feature selection algorithms to determine which features were the best in classifying red and white wines. Since there were 12 features in total, we decided to settle at selecting 6 features to determine the "better half" of all the features and see how they perform.

Model Exporation: Feature Selection Methods

Using the feature selection method, we wanted to find the most influential features and construct a model using those features that best describes the data. We used three types of feature selection: Univariate Selection, Recursive Feature Elimination and Principal Component Analysis. Using these selection methods, we generated three data sets.

Univariate Selection

```
[2016.51  4829.317  236.389  899.766 2315.829 1858.136 6252.796 1169.65
  789.05  2021.708    7.068   93.812]
```

Selected Features: fixed acidity, volatile acidity, chlorides, total sulfur dioxide

Recursive Feature Elimination

```
Num Features: 4
Selected Features: [False  True False False  True False False False  True  True False False]
Feature Ranking: [3 1 2 5 1 8 7 4 1 1 6 9]
```

Selected Features: volatile acidity, chlorides, pH, sulphates

Principal Component Analysis

```
Explained Variance: [9.536e-01 4.062e-02 4.826e-03 4.944e-04]
[[-7.408e-03 -1.184e-03  4.869e-04  4.102e-02 -1.682e-04  2.305e-01
   9.722e-01  1.772e-06 -6.555e-04 -7.043e-04 -5.452e-03 -5.327e-04]
 [-5.372e-03 -7.870e-04 -2.472e-04  1.863e-02  6.684e-05  9.726e-01
  -2.314e-01  1.278e-06  6.480e-04  3.465e-04  2.879e-03  9.152e-03]
 [ 2.385e-02  9.047e-04  1.922e-03  9.952e-01  1.766e-04 -2.713e-02
  -3.585e-02  4.608e-04 -6.911e-03 -1.936e-03 -8.260e-02 -8.792e-03]
 [ 7.134e-01  2.400e-02  2.403e-02 -7.050e-02  9.905e-03  1.081e-02
   2.261e-03  1.439e-03 -2.761e-02  2.236e-02 -6.098e-01 -3.341e-01]]
```

Selected Features: fixed acidity, volatile acidity, citric acid, residual sugar

Performance Validation

After we selected features, we wanted to find out their performance to select the best model. We used three classifiers: Random Forest, Neural Network and Decision Tree to test their performance. We split the data into training data and validation data and used Area Under the Curve (AUC) to evaluate the result.

Random Forest

```
Base Without Feature Selection
Training AUC: 0.9993839835728953
Validation AUC: 0.9961578668164512

With Unvariate Selection
Training AUC: 0.9996840941399464
Validation AUC: 0.995404142602701

With Recursive Feature Elimination
Training AUC: 0.9967556468172485
Validation AUC: 0.9885939202780619

With Principal Component Analysis
Training AUC: 0.994152582530406
Validation AUC: 0.9792004722378249
```

Neural Networks

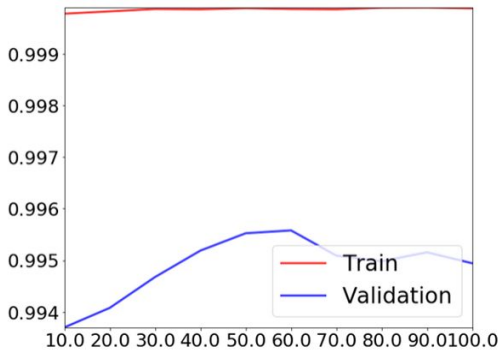
```
Base Without Feature Selection
Training AUC: 0.9958774285262991
Validation AUC: 0.9930107481712986

With Unvariate Selection
Training AUC: 0.9966956247038383
Validation AUC: 0.9925380702739811

With Recursive Feature Elimination
Training AUC: 0.9832696256515558
Validation AUC: 0.9818051014807241

With Principal Component Analysis
Training AUC: 0.9898531037750751
Validation AUC: 0.9787609998127693
```

Decision Trees

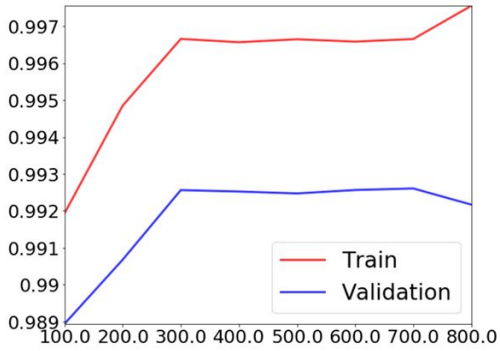
```
Base Without Feature Selection
Training AUC: 1.0
Validation AUC: 0.9764777235288377

With Unvariate Selection
Training AUC: 1.0
Validation AUC: 0.9640768798698025

With Recursive Feature Elimination
Training AUC: 0.9999810456483967
Validation AUC: 0.921591078209689

With Principal Component Analysis
Training AUC: 0.9993460748696887
Validation AUC: 0.9234257805686378
```

All these three classifiers indicate that Univariate Selection generates a better dataset than the other two feature selection methods. Hence, we decided to focus on the Univariate Selection method. In order to improve the performance of our classifier, we analyzed different parameters to find the best configuration. Here are some parameters we tried:

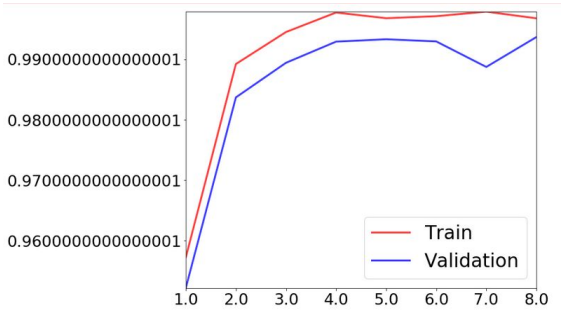
Estimator of Random Forest Classifier (10 - 100):



Number of iterations of Neural Network Classifiers (100 - 800)



We also tried to find out what was the most optimal number of features to be selected and if any were unnecessary. By comparing the performance of the classifier on different numbers of features, we find that performance is the best around 4 to 7.



Reflection on Models: Adapting to Under and Overfitting

Overall, the classifier shows an interesting result that fixed acidity, volatile acidity, chlorides, total sulfur dioxide are four features that contain valuable information about the type of wine. PCA yielded the worst results out of the three feature selection algorithms, performing worse in almost all scenarios with different classifiers, the only exception being the recursive feature elimination on decision trees. We can see some cases of overfitting when using decision trees with recursive feature elimination and PCA. The validation scores are considerably more distant compared to all other algorithms, and although the difference might seem subtle, it is well above its competitors. The overall best feature selection algorithm is none, but at the cost of using more features. Lowering a significant portion of our features improved performance and the measured difference wasn't all that great.

Work Cited

Brownlee, Jason. “Feature Selection For Machine Learning in Python.” Machine Learning Mastery, 18 Dec. 2019, machinelearningmastery.com/feature-selection-machine-learning-python/.

“Learn.” *Scikit*, scikit-learn.org/stable/.

Narkhede, Sarang. “Understanding AUC - ROC Curve.” *Medium*, Towards Data Science, 26 May 2019, towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.