Modeling and Predicting Wine Quality: Machine Learning Approach

2024-05-02

# Modeling and predicting diabetes

## Introduction

The quality of the wine is influenced by a number of factors including but not limited to sensory characteristics and chemical composition. It is in the interest of the manufacture to determine what affects the quality of wine. In this study, we employ machine learning method to model and predict quality of the wine based on the physicochemical attributes. The data set used in this study is obtained from Kaggle website with variable including acidity, volatilize acidity, citric acid, residual sugar, chloride, free sulfuric dioxide, total sulfuric dioxide, density, pH, sulphates and alcohol content. The dependent variable for the study is the quality of the wine.

The application of machine learning in this study provides power tools analyzing and modeling wine quality. By doing so, this paper aims at evaluating the relationship between chemical composition and their effect on the quality of the wine. This algorithm will help determine the most appropriate machine learning technique that can help classify and model wine according their respective quality categories. In this study, wine quality will be categorized into three main categories, that is, “bad”, “average”, and “good”.

## Objectives

This study is guided by the following objectives

* To evaluate the performance of various machine learning algorithms, which include Naive Bayes, k-Nearest Neighbors (kNN), Hierachical clustering and K-Means Clustering in modeling classification of wine quality
* Assessing the effectiveness of the developed models in classifying and predicting the quality of the wine.

## Methodology

This study employed the use of secondary data (Red Wine Data) obtained from Kaggle website. (<http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv>) The dataset used comprised the chemical characteristics of the wine and the quality of the wine. The machine learning applied in this study include the following;

* Classification and Regression Tree (CART): CART is a decision tree algorithm that recursively splits the data into subsets based on the value of predictor variables. At each step, it chooses the variable that best splits the data, resulting in a tree-like structure where the leaves represent the predicted outcome.
* Random Forest: Random Forest is an ensemble machine learning method constructed from various decision trees to create one classification and prediction algorithm.
* k-Nearest Neighbors: This algorithm is a non-parametric machine learning algorithm thta classifies an individual based on the k-Nearest neighbors.
* Support Vector Machine (SVM): SVM is a supervised learning algorithm used for classification and regression tasks. It works by finding the hyper-plane that best separates the classes in the feature space. Using this algorithm, three kernels options are always specified, that is Sigmoid, Linear and Polynomial, however, in many instance, linear kernel has always outperformed the sigmoid and polynomial kernel.
* Naive Bayes Classifier: This is based on Bayes’ theorem to classify individual, holding the assumption that feature are independent

## 

## Results

### Summary Statistics

|  | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fixed.acidity | 1 | 4898 | 6.85 | 0.84 | 6.80 | 6.82 | 0.74 | 3.80 | 14.20 | 10.40 | 0.65 | 2.17 | 0.01 |
| volatile.acidity | 2 | 4898 | 0.28 | 0.10 | 0.26 | 0.27 | 0.09 | 0.08 | 1.10 | 1.02 | 1.58 | 5.08 | 0.00 |
| citric.acid | 3 | 4898 | 0.33 | 0.12 | 0.32 | 0.33 | 0.09 | 0.00 | 1.66 | 1.66 | 1.28 | 6.16 | 0.00 |
| residual.sugar | 4 | 4898 | 6.39 | 5.07 | 5.20 | 5.80 | 5.34 | 0.60 | 65.80 | 65.20 | 1.08 | 3.46 | 0.07 |
| chlorides | 5 | 4898 | 0.05 | 0.02 | 0.04 | 0.04 | 0.01 | 0.01 | 0.35 | 0.34 | 5.02 | 37.51 | 0.00 |
| free.sulfur.dioxide | 6 | 4898 | 35.31 | 17.01 | 34.00 | 34.36 | 16.31 | 2.00 | 289.00 | 287.00 | 1.41 | 11.45 | 0.24 |
| total.sulfur.dioxide | 7 | 4898 | 138.36 | 42.50 | 134.00 | 136.96 | 43.00 | 9.00 | 440.00 | 431.00 | 0.39 | 0.57 | 0.61 |
| density | 8 | 4898 | 0.99 | 0.00 | 0.99 | 0.99 | 0.00 | 0.99 | 1.04 | 0.05 | 0.98 | 9.78 | 0.00 |
| pH | 9 | 4898 | 3.19 | 0.15 | 3.18 | 3.18 | 0.15 | 2.72 | 3.82 | 1.10 | 0.46 | 0.53 | 0.00 |
| sulphates | 10 | 4898 | 0.49 | 0.11 | 0.47 | 0.48 | 0.10 | 0.22 | 1.08 | 0.86 | 0.98 | 1.59 | 0.00 |
| alcohol | 11 | 4898 | 10.51 | 1.23 | 10.40 | 10.43 | 1.48 | 8.00 | 14.20 | 6.20 | 0.49 | -0.70 | 0.02 |
| quality | 12 | 4898 | 5.88 | 0.89 | 6.00 | 5.85 | 1.48 | 3.00 | 9.00 | 6.00 | 0.16 | 0.21 | 0.01 |

The mean fixed acidity was 6.85 (SD = 0.84), volatile acidity was 0.28 (SD = 0.10), citric acid was 0.33 (SD = 0.12), residual sugar was 6.39 (SD = 5.07), chlorides were 0.05 (SD = 0.02), free sulfur dioxide was 35.31 (SD = 17.01), total sulfur dioxide was 138.36 (SD = 42.50), density was 0.99 (SD = 0.00), pH was 3.19 (SD = 0.15), sulphates were 0.49 (SD = 0.11), alcohol content was 10.51 (SD = 1.23), and wine quality was 5.88 (SD = 0.89).

**Model One: Classification and Regression Trees (CART) Model**

|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 0.75330 |
| Kappa | 0.45420 |
| Sensitivity |  |
| bad | 0.00000 |
| average | 0.90090 |
| good | 0.56960 |
| Specificity |  |
| bad | 1.00000 |
| average | 0.51140 |
| good | 0.90500 |

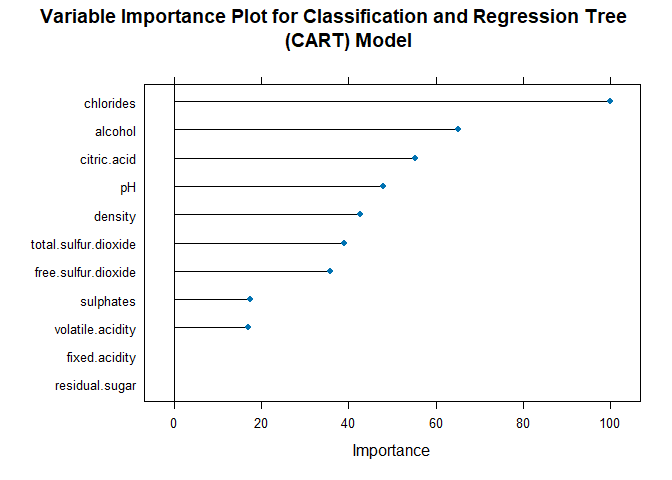
The classification and regression model estimated shows that the model has an accuracy of approximately 75.33%. This shows that the model correctly classifies the wine qualities into their respective qualities (bad, average and good), 75.33% of the time.

### Variable Importance

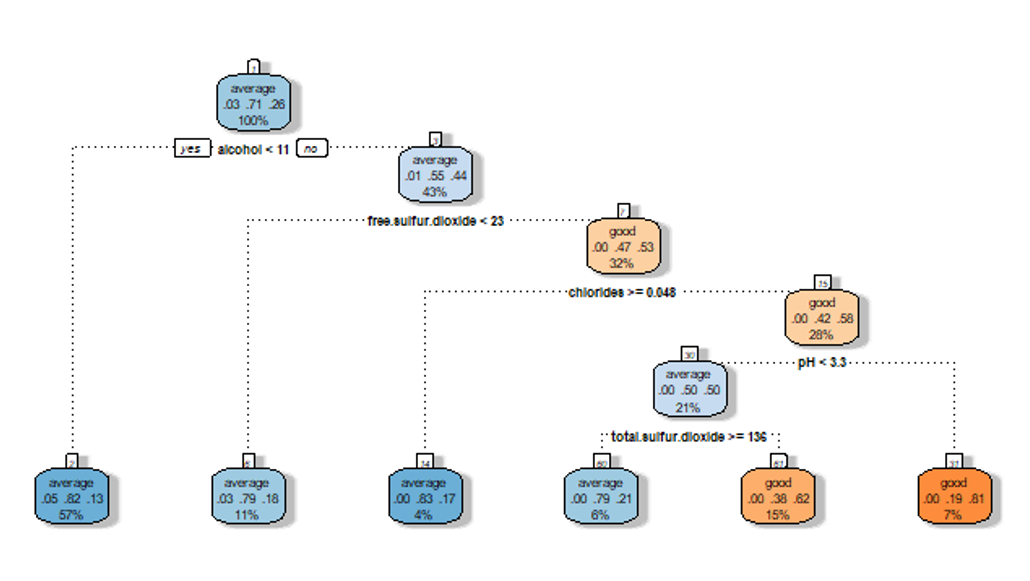
|  |  |
| --- | --- |
| Variables | Overall Importance (%) |
| alcohol | 100.0000 |
| density | 80.6661 |
| chlorides | 53.9612 |
| Total sulfur dioxide | 40.1294 |
| sulphates | 21.7334 |
| Free sulfur dioxide | 0.0000 |
| pH | 0.0000 |
| Citric acid | 0.0000 |
| Volatile acidity | 0.0000 |
| Fixed acidity | 0.0000 |

The results shows that the most important and significant variable in the classification and regression trees model developed are alcohol, chlorides, density, total sulfuric dioxide and residual sugar ans so on. From the results, alcohol has 100% importance, followed by density with 80.67%, chlorides with 53.96% and total sulfuric dioxide with 40.13%%. Consider the plot below to visualize the results above.

### Plot the Variable Importance



### Plot the Classification and Regression Tree Model



The plot shows the classification and regression tree diagram for classifying wine qualities into bad, average and good quality wine.

## Model Two: Random Forest

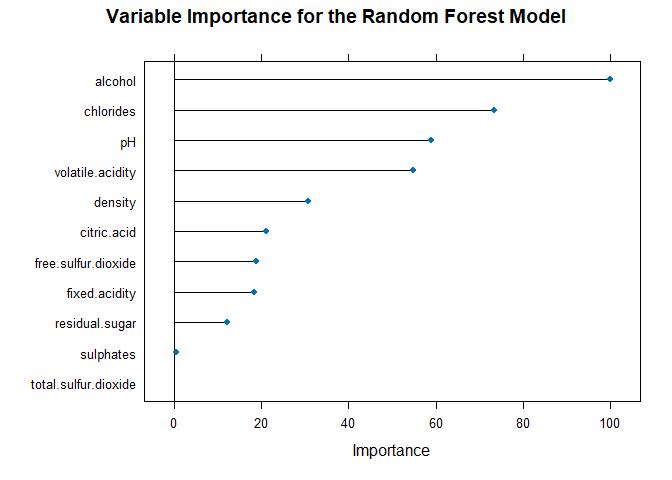
|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 1.00000 |
| Kappa | 1.00000 |
| Sensitivity |  |
| bad | 1.00000 |
| average | 1.00000 |
| good | 1.00000 |
| Specificity |  |
| bad | 1.00000 |
| average | 1.00000 |
| good | 1.00000 |

The random forest model developed shows that the model has an accuracy of approximately 100%. This shows that the model classifies wine qualities correctly into their respective wine qualities as either bad, average, or good, 100% of the time.

### Obtain variable importance

|  |  |
| --- | --- |
| Variables | Overall |
| alcohol | 100 |
| density | 59.379 |
| chlorides | 43.362 |
| Total sulfur dioxide | 38.624 |
| Free sulfur dioxide | 27.064 |
| sulphates | 19.342 |
| Residual sugar | 18.279 |
| Volatile acidity | 16.669 |
| Citric acid | 12.273 |
| pH | 4.8563 |

The results above show relative importance of each variable in helping predicting and classify wine qualities. From the results, alcohol has a higher relative importance of 100%, followed by density with 59.38%, chlorides with 43.36%, and total sulphuric with dioxide with 38.62% and so so. The figure below shows the relative importance of each variable in our model



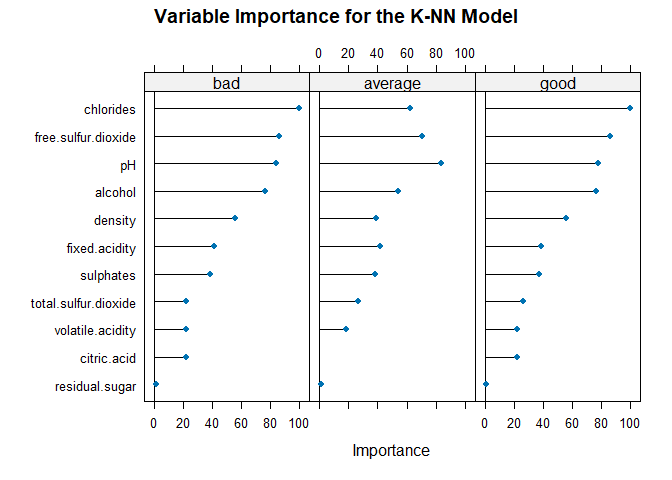
## Model Three: k-Nearest Neighbors

|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 0.77330 |
| Kappa | 0.44730 |
| Sensitivity |  |
| bad | 0.00000 |
| average | 0.86320 |
| good | 0.62030 |
| Specificity |  |
| bad | 1.00000 |
| average | 0.56820 |
| good | 0.86430 |

k-Nearest Neighbors performed slightly better classification and regression tree. However, the model performs slightly poorer in prediction as compared to random forest. From the above algorithm, the k-Nearest Neighbors accuracy is approximately 80% implying that the model correctly predicts and classify wine quality in their correct wine qualities 80% of the time. The algorithm has a higher mis-classification error than that of random forest and a lower mis-classification error as compared to classification and regression tress (CART).

|  |  |  |  |
| --- | --- | --- | --- |
| bad | bad | average | good |
| Free sulfur dioxide | 100 | 100 | 98.33204 |
| pH | 69.12381 | 52.28344 | 69.12381 |
| alcohol | 54.32323 | 65.35274 | 65.35274 |
| chlorides | 59.95531 | 54.31826 | 59.95531 |
| density | 35.06937 | 50.66327 | 50.66327 |
| Fixed acidity | 49.608 | 37.01232 | 49.608 |
| Total sulfur dioxide | 37.91862 | 38.73491 | 38.73491 |
| Citric acid | 33.36665 | 11.9532 | 33.36665 |
| Volatile acidity | 31.53295 | 23.96164 | 31.53295 |
| Residual sugar | 29.12753 | 29.12753 | 15.94649 |

The relative importance of each variable in the classification and prediction of wine quality is as shown above, with free sulfuric dioxide as the overall important variable in the classification and prediction of bad, average and good wine, followed by pH in classifying bad wine, alcohol in classifying average wine quality and pH in classifying good quality wine. This is also indicated in the plot below



## Model Four: Naive Bayes

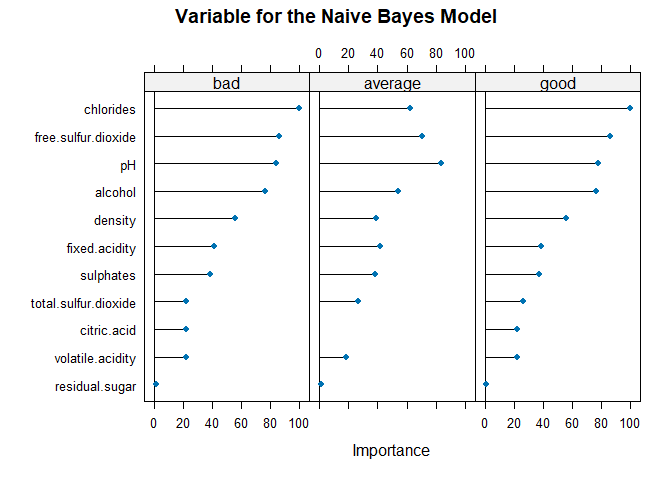
|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 0.75330 |
| Kappa | 0.42360 |
| Sensitivity |  |
| bad | 0.33333 |
| average | 0.82550 |
| good | 0.62760 |
| Specificity |  |
| bad | 0.99676 |
| average | 0.59090 |
| good | 0.83260 |

Similar to the k-nearest neighbors, naive bayes performs slightly poor in the classification and prediction of the wine quality. From the model above, naive bayes correctly predict and classify wine qualities in their respective wine quality categories as either bad, average and good, 77% of the time, which lower as compared to random forest and above that of CART model.

### Variable Importance

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | bad | average | good |
| Free sulfur dioxide | 100 | 100 | 98.332 |
| pH | 69.124 | 52.283 | 69.124 |
| alcohol | 54.323 | 65.353 | 65.353 |
| chlorides | 59.955 | 54.318 | 59.955 |
| density | 35.069 | 50.663 | 50.663 |
| Fixed acidity | 49.608 | 37.012 | 49.608 |
| Total sulfur dioxide | 37.919 | 38.735 | 38.735 |
| Citric acid | 33.367 | 11.953 | 33.367 |
| Volatile acidity | 31.533 | 23.962 | 31.533 |
| Residual sugar | 29.128 | 29.128 | 15.946 |

The relative importance of each variable in the classification and prediction of wine quality is as shown above, with free sulfuric dioxide as the overall important variable in the classification and prediction of bad 100%), average (100%) and good (100%) wine, followed by pH in classifying bad wine (69.12%), alcohol in classifying average (65.35%) wine quality and pH in classifying good quality wine (69.12%). This is also indicated in the plot below



## Model Five: Support Vector Machine (SVM)

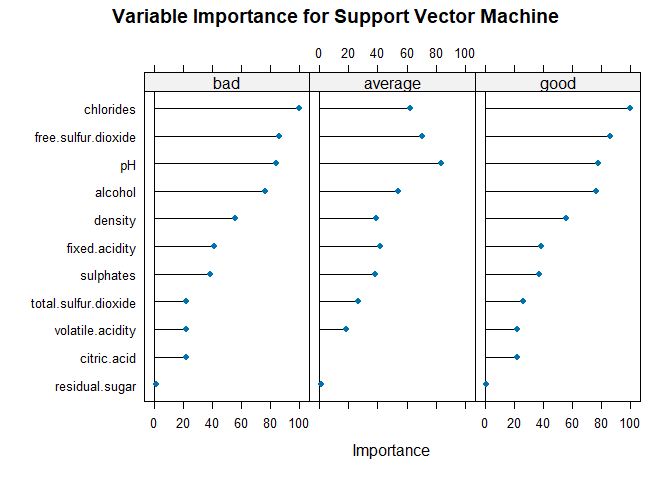
|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 0.75330 |
| Kappa | 0.00000 |
| Sensitivity |  |
| bad | 0.00000 |
| average | 1.00000 |
| good | 1.00000 |
| Specificity |  |
| bad | 1.00000 |
| average | 0.00000 |
| good | 1.00000 |

The support vector machine model estimated above shows that the model has an accuracy of approximately 75.33%. This shows that the model correctly classifies the wine qualities into their respective qualities (bad, average and good), 78.67% of the time with a relatively higher mis-classification error of approximately 24.67%

### Variable Importance

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | bad | average | good |
| Free sulfur dioxide | 100 | 100 | 98.332 |
| pH | 69.124 | 52.283 | 69.124 |
| alcohol | 54.323 | 65.353 | 65.353 |
| chlorides | 59.955 | 54.318 | 59.955 |
| density | 35.069 | 50.663 | 50.663 |
| Fixed acidity | 49.608 | 37.012 | 49.608 |
| Total sulfur dioxide | 37.919 | 38.735 | 38.735 |
| Citric acid | 33.367 | 11.953 | 33.367 |
| Volatile acidity | 31.533 | 23.962 | 31.533 |
| Residual sugar | 29.128 | 29.128 | 15.946 |

free sulfuric dioxide has 100% importance in classifying and predicting bad quality wine and 100% importance in classifying and predicting average quality wine and finally 98.33% importance in classifying and predicting good quality. On the other hand, pH has 69.12% importance in classifying bad quality wine, 52% importance in classifying average quality wine and 60.12% importance in classifying good quality wine. The remaining variables and their importance in classifying wine qualities into their respective wine quality categories is as shown in the table above. Besides, the results can be visualized as shown below.



**Comparison for the Overall Best Machine Learning Algorithms**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **Support Vector Machine** | **Naïve Bayes** | **K-Nearest Neighbors** | **Random Forest** | **CART** |
| **Accuracy** | 0.75330 | 0.75330 | 0.77330 | 1.00000 | 0.75330 |
| **Kappa** | 0.00000 | 0.42360 | 0.44730 | 1.00000 | 0.45420 |
| **Sensitivity** |  |  |  |  |  |
| bad | 0.00000 | 0.33333 | 0.00000 | 1.00000 | 0.00000 |
| average | 1.00000 | 0.82550 | 0.86320 | 1.00000 | 0.90090 |
| good | 1.00000 | 0.62760 | 0.62030 | 1.00000 | 0.56960 |
| **Specificity** |  |  |  |  |  |
| bad | 1.00000 | 0.99676 | 1.00000 | 1.00000 | 1.00000 |
| average | 0.00000 | 0.59090 | 0.56820 | 1.00000 | 0.51140 |
| good | 1.00000 | 0.83260 | 0.86430 | 1.00000 | 0.90500 |

In order to determine the overall best model among the machine learning algorithms, we can consider various performance metrics such as accuracy, kappa, sensitivity, and specificity across different classes. From the results above, Random Forest has the highest accuracy of 1.00000, indicating it correctly classified all instances. Besides, the algorithm has also has the highest kappa of 1.00000, indicating perfect agreement beyond chances. For the "bad" class, Random Forest and K-Nearest Neighbors have the highest sensitivity of 1.00000, indicating they correctly identified all instances of the "bad" class. For the "average" class, Random Forest and CART have the highest sensitivity of 1.00000 and for the "good" class, Random Forest has the highest sensitivity of 1.00000. On the other hand, the model has a specificity of 1.0000 for the bad wine quality. It is important to note that for the "average" class, Naïve Bayes and Random Forest have the highest specificity of 1.00000. However, for the "good" wine quality, Random Forest has the highest specificity of 1.00000. Considering all these metrics, Random Forest is the overall best model as it achieves the highest accuracy, kappa, sensitivity, and specificity across different classes of wine quality (bad, average and good).