LeTicia Cancel

DATA622 Homework #3

December 12, 2022

I work in the admissions office at a college here in New York so I looked for decision tree data modeling in higher education. A study conducted by Ying-Sing Liu & Liza Lee on [nature.com](https://www.nature.com/articles/s41599-022-01413-z) used decision trees to evaluate college admissions1. Their study looked at 5-year junior college enrollment to explore why some students fail the admission process in Taiwan and where the admission process can be improved. The decision tree allowed them to trace the different paths where applicants failed to enroll and which path led to the largest number of failures.

Another study conducted by Vasiliki Matzavela and Efthimios Alepis on sciencedirect.com used decision tree modeling to predict the students' performance2. The decision tree allowed them to see the characteristics of students who had excellent grades and students who had lower grades.

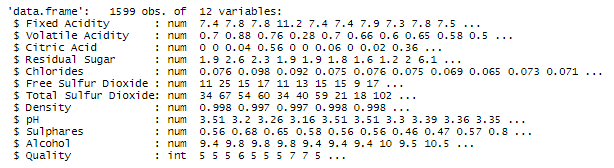
The dataset from the previous assignment was used for this analysis. The dataset used for this assignment is the [Wine Quality Data Set](https://archive.ics.uci.edu/ml/datasets/wine+quality) from the UCI Machine Learning Repository. The datasets available are Red Wine Quality and White Wine Quality. The Decision Tree was built with the Red Wine dataset. The dataset has 1,599 observations and 12 variables. The variables are:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol

Output variable (based on sensory data):

1. quality (score between 0 and 10)

If we examine the structure of the dataframe, we see that the datatype of 11 out of the 12 variables are numerical and the output variable 'Quality' is an integer.

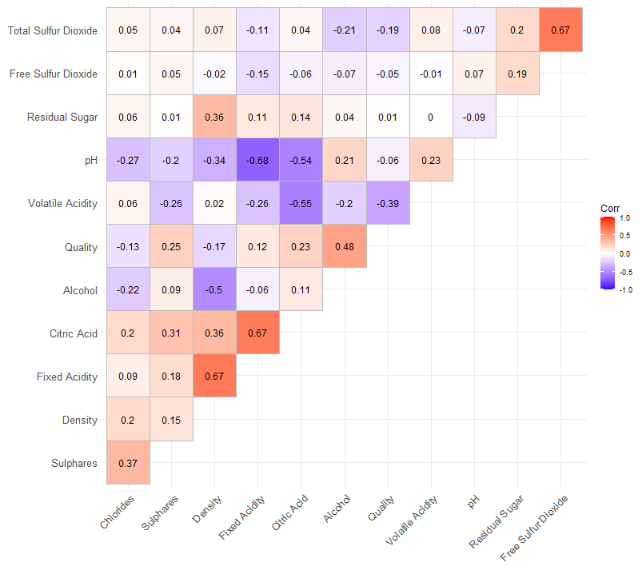


Since most of the variables are numbers we can check if any are correlated. Residual Sugar and Volatile Acidity have a correlation coefficient of zero so there is zero correlation between these two variables. The top negatively correlated variables are:

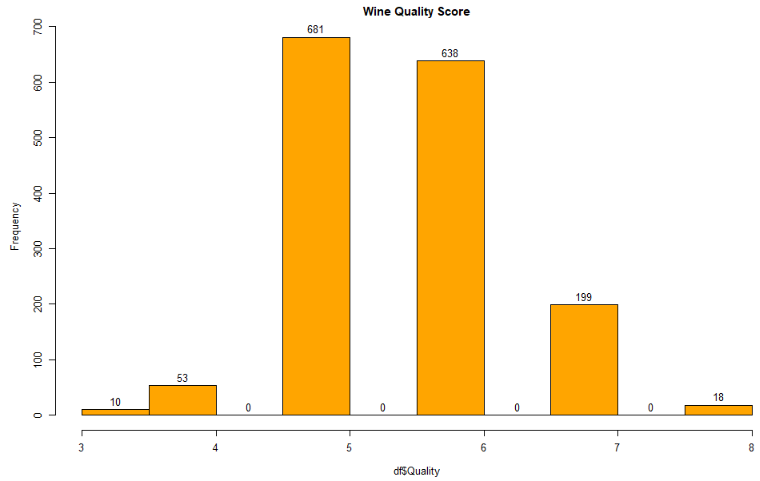
* pH and Fixed Acidity -0.68
* Volatile Acidity and Citric Acid -0.55
* pH and Citric Acid -0.54
* Alcohol and Density -0.5

The top positively correlated variables are:

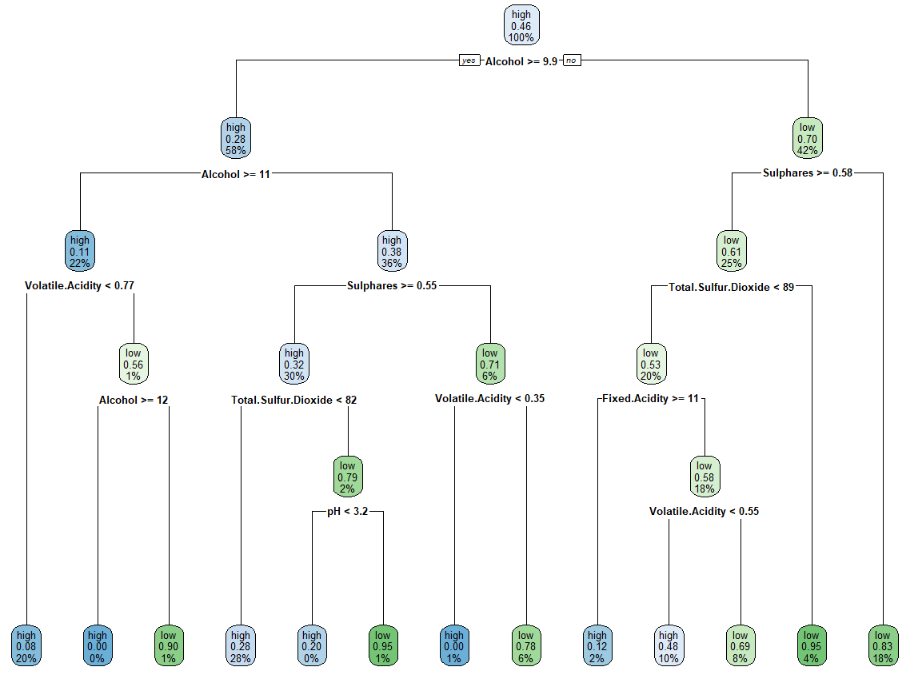
* Three variable pairs are tied for the top positive correlation coefficient of 0.67
* Total Sulfur Dioxide and Free Sulfur Dioxide
* Citric Acid and Fixed Acidity
* Density and Fixed Acidity
* Quality and Alcohol 0.48



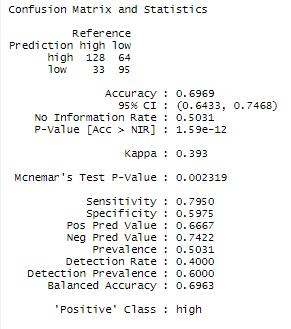
The Quality score will be used as the predictor for this analysis. The Quality score has a range between 0 and 10 with zero being the lowest score and 10 being the highest score. High = 855/53.47%, low = 744/46.53%, 1,599 total. Over 50% have a score of 6 or above. We will create a new column and group scores 6+ as "High" and all others as "Low".



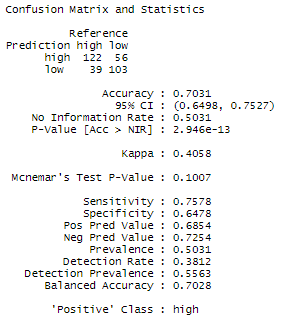
A new column 'qualityScore' is added to the dataframe. All rows with a score of 6 or above is labeled as High and all scores below 6 are labeled as Low. The data is split 80/20 into test and training datasets. A Decision Tree is then created using the qualityScore as the predictor.



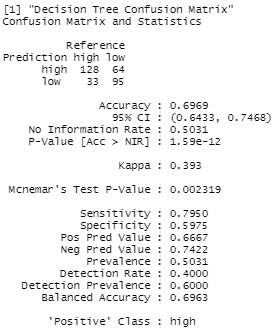
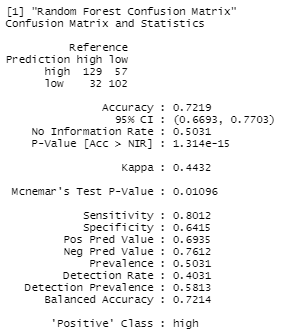
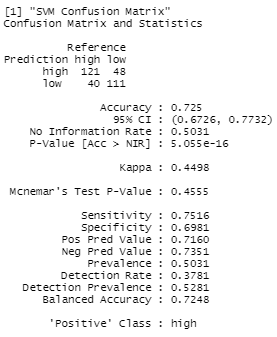
A confusion matrix is created with the predictions from the test dataset. The accuracy score is a little low at 69.69% but we will create a Random Forest and compare the results.



The accuracy of the Decision Tree is higher at 70.31%



The SVM model was used on the same dataset with a confusion matrix so we can evaluate which is best for this dataset. If we compare all three confusion matrix results, we see that the accuracy is similar across all three models but the SVM model is the highest at 72.5%. The SVM model also has the lowest sensitivity and the highest specificity. The p-value is lowest with the Decision Tree at 0.002319. If I had to choose one model based on the confusion matrix results, I would choose the Random Forest because the results look more balanced compared to the other models. The Decision Tree has the lowest p-value but also has the lowest accuracy and specificity. The SVM model has the highest accuracy but the highest p-value. The Random Forest has the second highest accuracy but the score is only 0.0031 from the SVM model.



References

1. Liu, Y.-S., &amp; Lee, L. (2022, October 25). Evaluation of College Admissions: A decision tree guide to provide information for improvement. Nature News. Retrieved December 12, 2022, from <https://www.nature.com/articles/s41599-022-01413-z>

2. panelVasilikiMatzavelaPersonEnvelopeEfthimiosAlepisEnvelope, A. links open overlay, VasilikiMatzavelaPersonEnvelope, EfthimiosAlepisEnvelope, &amp; AbstractIn the area of machine learning and data science. (2021, October 5). Decision tree learning through a predictive model for student academic performance in Intelligent M-Learning Environments. Computers and Education: Artificial Intelligence. Retrieved December 12, 2022, from <https://www.sciencedirect.com/science/article/pii/S2666920X21000291>