

**Suitability of Confusion Matrices on the Quality of Wine & it's Attributes**​



**FINAL PROJECT**

**MBA 540.01 DATA MINING**

**BY: STEVEN QUIRINO, ZACHARY SANFORD & CAMERON LUCAS**​

3. Final Report

Your report needs to have the following sections:

• Abstract: Summarize the major goal of your paper, the techniques used in the

paper and the key results. (up to 300 words).

* This report determines the suitability of different confusion matrices, allowing us to visualize the performance of an algorithm and model for data regarding wine quality using python. These models include K-nearest neighbor (KNN), Support Vector Machines (SVM) and Random Forest (Decision Tree). Each model has different results that show the best model for wine quality predictions. Adjusting the attributes to get better results and understanding of what attributes are most important in the data.

• Introduction: Discuss the following items in the introduction:

o The background of the project and specify your problem.

* Wine has been tested by users of varying attributes. These users then give a number score to represent the quality of each wine. The data we were given includes different attributes along with the score. These included fixed acidity, citric acid, residual sugar, chlorides etc. These attributes are to be used to figure out what is most important when determining the quality of wine.

o The importance or practical value of the problem

* In many situations there are attributes that contribute to the quality of a product. This problem at hand can be an example on how to determine the best attributes of any consumer product, given similar data. If given the same attributes of any wine we could use the best performing model to make a data based prediction on whether the wine would be considered high or low quality.

• Methodology: Discuss the methodology you use.

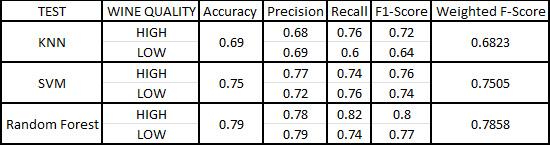
* Determine the models that best forecast wine quality, then re-testing by removing attributes to see which attributes the data are most reliant on.

• Data and Experimental settings: Discuss the data, evaluation metrics, and

anything else related to the experimental settings.

* The raw data lacked a qualitative aspect, so I needed to add it.​ The data contained a quantitative score for the quality of each wine tested​. I used an IF statement {=IF(L2>5,"High","Low")}to create a qualitative label, if the quality was above a 5 it would be considered high and if it was below 5 it would be considered low​. Creating this qualitative attribute allowed us to determine features and class parts (X & Y)​. First, I imported the necessary packages and data​. With any machine learning algorithm, it is important to divide datasets into train and test sets. Using train\_test\_split function​. I prepared the results for K-nearest neighbor, SVM and Random Forest. The results all differed based on the precision and accuracy of each test. Suitability was determined by these two metrics for the entire original dataset.

• Results and analysis: Show your results in figures and/or tables. Discuss them.



* We can see in K-nearest neighbor that there is a low accuracy but a good spread on the precision. When looking at Support Vector Machines there is a worse precision spread than KNN but the accuracy is higher and therefore better. The precision spread for Random Forest is the same as KNN and better than SVM but it is apparent that decision trees are more suitable based on the accuracy.
* The main way I evaluated the suitability of each confusion matrix was by calculating the F1-Score weighted average. F1-score is the harmonic mean of precision and recall. This is another metric to measure the accuracy of a model. This is a good metric for binary classification systems such as the one presented in this data​. Taking the number of high and low quality wines relative to their score. I did this by using the following formula:

weighted average of F1-Score = (“High”F1-Score) \* W\_high + (“Low”F1-Score) \* W\_Low

W\_high & W\_Low are calculated by taking each high and low tested and dividing it by the entire test data set. As you can see above Random forest is the most suitable for data predictions on wine.

* The weighted average of F1-Score for KNN on high- and low-quality wine is ​

= (.72 \* .5291) + (.64 \* .4708) = .6823​

* The weighted average of F1-Score for SVM on high- and low-quality wine is​

= (.76 \* .5291) + (.74 \* .4708) = .7505

* The weighted average of F1-Score for Random Forest on high- and low-quality wine is​

= (.80 \* .5291) + (.77 \* .4708) = .7858​

• Discussion and Conclusion: provide a conclusion of your project and discuss

future work.

* The results of the weighted average F1-Score points towards Random Forest being the most suitable. This test demonstrated that Random Forest had the best test results when the results were compared to each model. In order to be confident in the predictions of a model, a data scientist would want the highest precision, accuracy and weighted F1-Score possible. That would require retesting and changes to the data set to shoot for higher scores.