COSC2670: Practical Data Science

Assignment 2: Modelling wine quality based on physicochemical tests

Due midday on Monday, 22 May 2017

Submitted by:

Casey-Ann Charlesworth (3132392)

Table of Contents

[1. Executive summary 1](#_Toc482543192)

[2. Introduction 1](#_Toc482543193)

[3. Methodology 2](#_Toc482543194)

[4. Results 4](#_Toc482543195)

[4.1 Exploration 4](#_Toc482543196)

[4.2 Modelling 8](#_Toc482543197)

[5. Discussion 11](#_Toc482543198)

[6. Conclusion 11](#_Toc482543199)

[7. References 12](#_Toc482543200)

[8. Appendix 13](#_Toc482543201)

Table of Tables

[Table 1: Value counts of each quality rating (both red and white wines) 2](#_Toc482543202)

[Table 2: Summary statistics for red wine 13](#_Toc482543203)

[Table 3: Summary statistics for white wine 13](#_Toc482543204)

Table of Figures

[Figure 1: Proportion of wine quality rating broken down by dataset 3](#_Toc482543205)

[Figure 2: Boxplot of all variables (red wine) 4](#_Toc482543206)

[Figure 3: Boxplot of all variables (white wine) 5](#_Toc482543207)

[Figure 4: Histogram of all variables (red wine) 5](#_Toc482543208)

[Figure 5: Histogram of all variables (white wine) 6](#_Toc482543209)

[Figure 6: Free v total sulfur dioxide (red wine) 6](#_Toc482543210)

[Figure 7: Free v total sulfur dioxide (white wine) 6](#_Toc482543211)

[Figure 8: Fixed v volatile acidity (red wine) 7](#_Toc482543212)

[Figure 9: Fixed v volatile acidity (white wine) 7](#_Toc482543213)

[Figure 10: Citric acid v residual sugar (red wine) 7](#_Toc482543214)

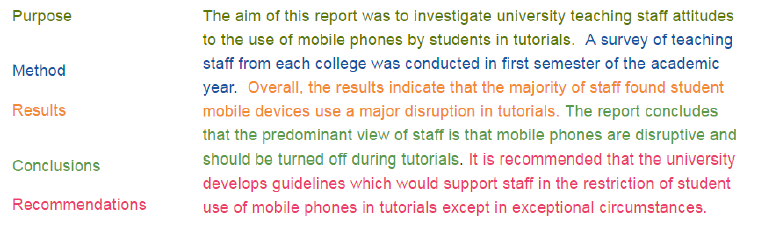
[Figure 11: Citric acid v residual sugar (white wine) 7](#_Toc482543215)

# Executive summary

The aim of this report was to investigate whether wine quality (both red and white varieties) could be modelled based on a number of physicochemical tests. The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. [1][2] Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). Overall, due to the unbalanced nature of the class variable (there were many more “average” wines compared to “bad” or “excellent” wines), the classification models employed (Decision Tree, K Nearest Neighbour, and Naïve Bayes) struggled to model and predict the quality of wine with greater than 60% accuracy.

[Still need conclusions and recommendations to be added]

*Guidance: A paragraph-length summary of the key arguments and findings*



# Introduction

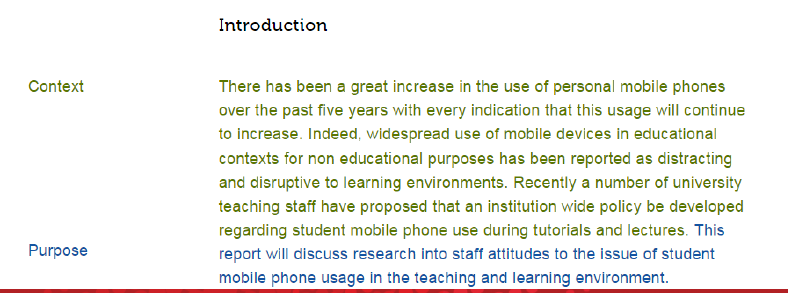
Wine quality is subjective. Personal preferences and sensory physiology of each individual can lead to one person’s preferences of wine being different to another’s. However, any wine drinker would surely welcome a little guidance in which brand and/or type may be considered better quality over another when it comes to guiding their purchases, whether or not price range is factor. But are there factors that can help or hinder the measure of wine quality, and do any individual factors dominate in determining whether a wine should be considered excellent (or not so excellent)?

This report will discuss whether physicochemical levels of a number of variables present in wine production can reveal a preference for some brands/types over another, and thus lead to a higher quality rating.

Complete

*Guidance:*

* *Explanation of the problem*
* *Particularly important since many readers might not be experts in the topic area, or the analytical methods that were applied*
* *Often includes a literature review*
* *Explain what’s already known, as well as gaps in knowledge*



# Methodology

Two datasets, being red wine and white wine were obtained from the UCI Machine Learning Repository [2]. These datasets were supplied by researchers who completed a similar paper entitled *Modeling wine preferences by data mining from physicochemical properties*. [1]

The datasets had identical variables, however, the red wine dataset contained 1599 observations while the white wine dataset contained 4898 observations. A breakdown of the summary statistics of both datasets can be found in section 8 (Appendix).

The physicochemical variables included the following measurements:

* fixed acidity (g(tartaric acid)/dm3)
* volatile acidity (g(acetic acid)/dm3)
* citric acid (g/dm3)
* residual sugar (g/dm3)
* chlorides (g(sodium chloride)/dm3)
* free sulfur dioxide (mg/dm3)
* total sulfur dioxide (mg/dm3)
* density (g/cm3)
* pH
* sulphates (g(potassium sulphate)/dm3)
* alcohol (% vol.)

The quality variable (being the designated “class” variable) was a sensory rating between 0-10.

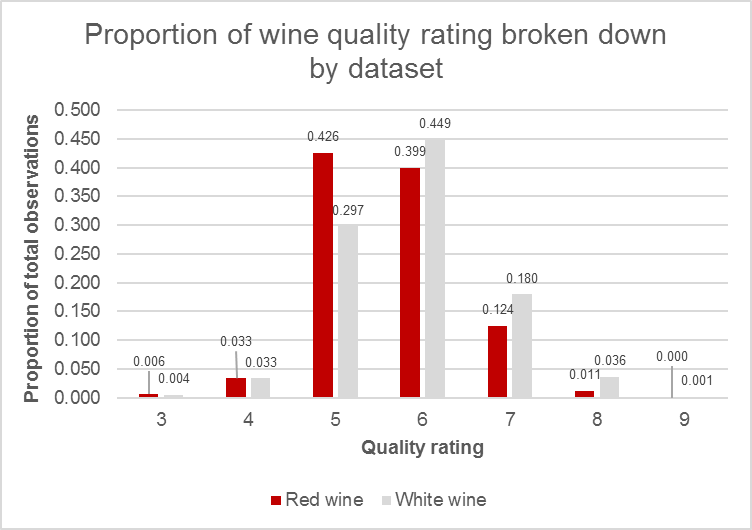
Both datasets were unbalanced, with the value counts of each quality rating noted below in Table 1.

Table 1: Value counts of each quality rating (both red and white wines)

| Red wine | | White wine | |
| --- | --- | --- | --- |
| **Quality** | **Count** | **Quality** | **Count** |
| 3 | 10 | 3 | 20 |
| 4 | 53 | 4 | 163 |
| 5 | 681 | 5 | 1457 |
| 6 | 638 | 6 | 2198 |
| 7 | 199 | 7 | 880 |
| 8 | 18 | 8 | 175 |
|  |  | 9 | 5 |

The proportions of each quality rating are shown in Figure 1 below.

Figure 1: Proportion of wine quality rating broken down by dataset



The following analysis and modelling was completed using iPython and includes a number of classification models from *sklearn*, explained as follows:

* **K Nearest Neighbour**

K-Nearest Neighbors, or simply kNN, belongs to the class of instance-based learning, also known as lazy classifiers. It’s one of the simplest classification methods because the classification is done by just looking at the K closed examples in the training set … [3]

* **Decision Tree**

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. [4]

* **Naïve Bayes**

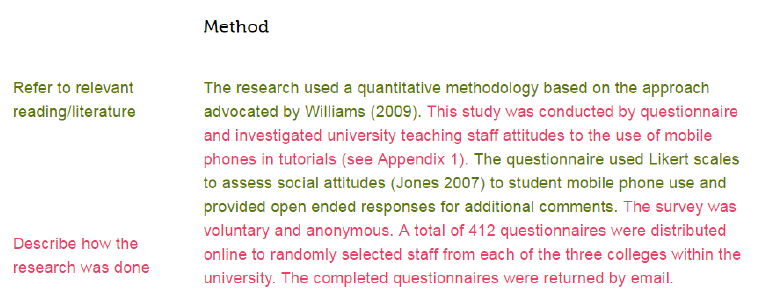
Naïve Bayes is a very common classifier used for probabilistic multiclass classification. Given the feature vector, it uses the Bayes rule to predict the probability of each class … it’s very effective with large and fat data (with many features) with a consistent a priori[[1]](#footnote-1) probability. [3]

It should be noted that although some outliers were detected, due to the dataset being previously cleaned by P. Cortez, et al [1], and the author of this paper not being knowledgeable in the area nor structure of the physicochemicals involved in the datasets, these outliers were retained for the analysis.

complete

*Guidance:*

* *Explanation of the research methods and data sources used*
  + *Data collection process (particularly if new data was gathered)*
  + *Choice of variables used for analysis*
  + *Analytical techniques and models used*



# Results

## Exploration

As this study looks to report on two discrete datasets, it was necessary to combine some functions. Therefore, each variable has been analysed below, for both datasets, however the analyses has been grouped into combined graphs. Both box plots and histograms were created for each variable per dataset in Figure 2 to Figure 5 below.

Figure 2: Boxplot of all variables (red wine)

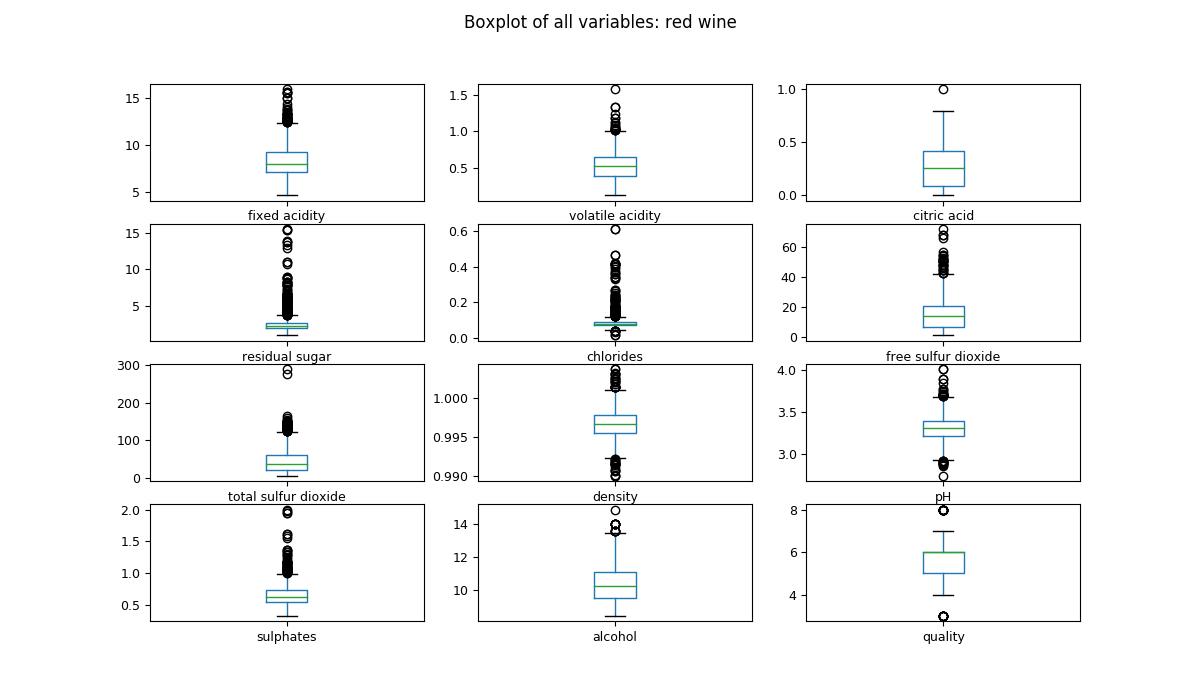
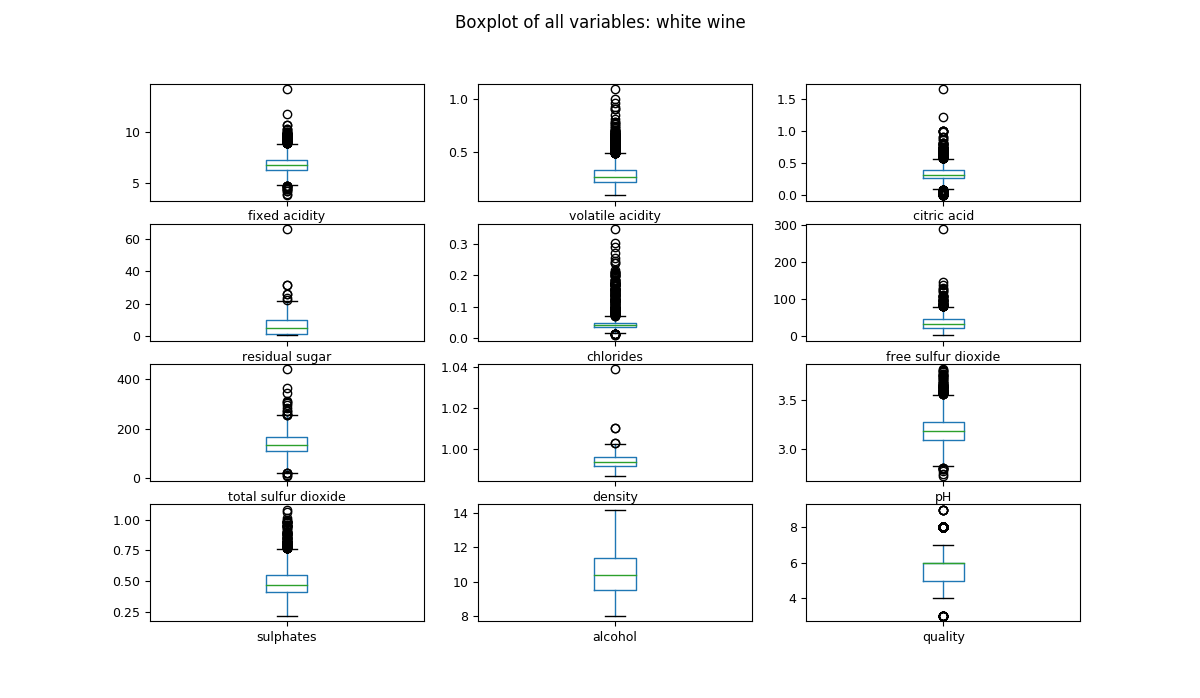


Figure 3: Boxplot of all variables (white wine)



A side by side comparison of the red and white boxplots reveals little difference (apart from outliers). Therefore, the combined histograms in Figure 4 and Figure 5 below should also be relatively similar in structure, with little variation between red and white wine.

Figure 4: Histogram of all variables (red wine)

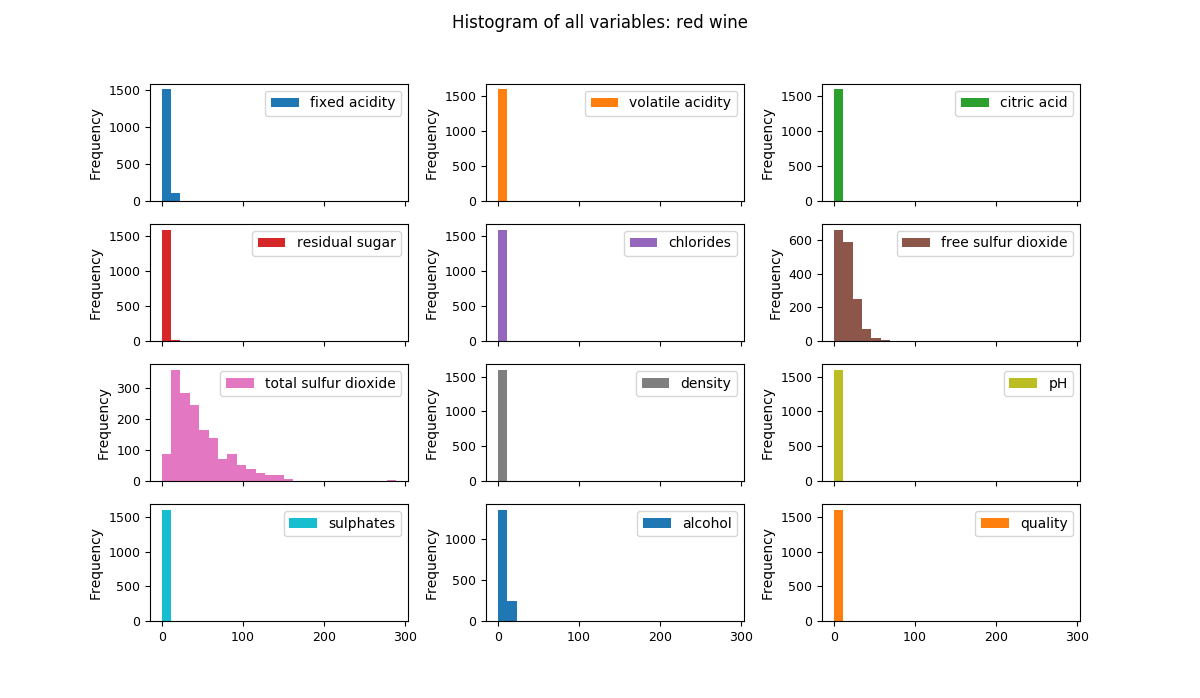
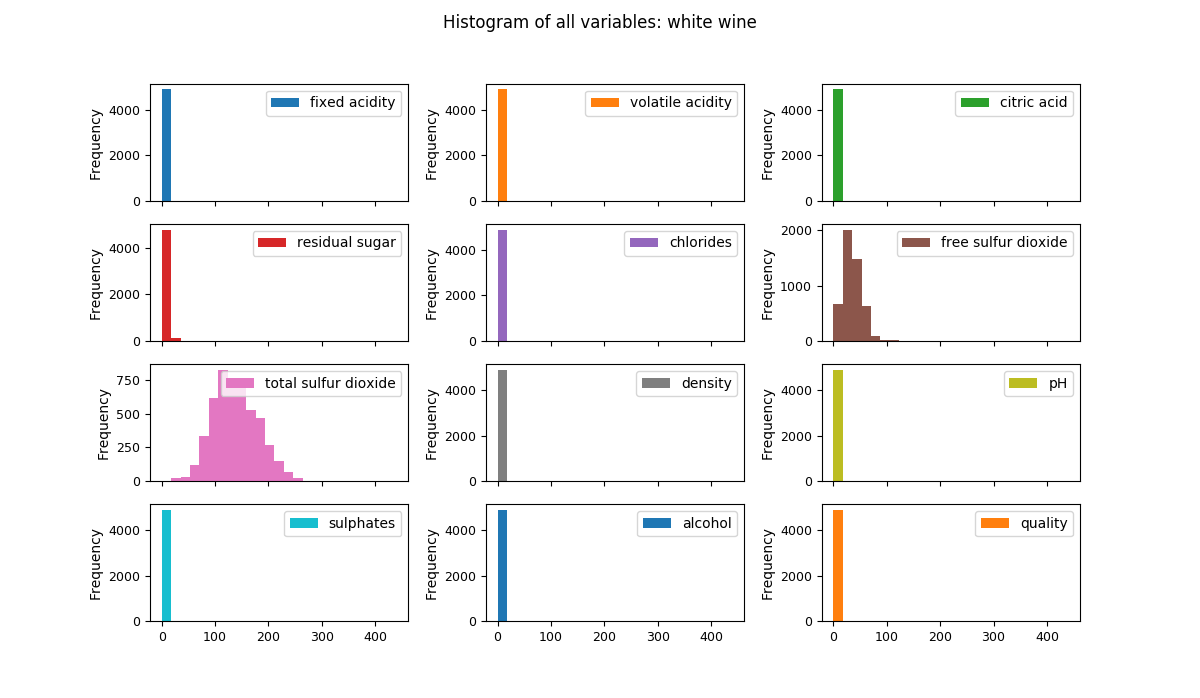


Figure 5: Histogram of all variables (white wine)



The only notable differences here were in the two sulfur dioxide variables.

To drill down further, the variables were split into pairs for analysis. This time, red and white graphs for the same pairs will be side by side for easy reference.

|  |  |
| --- | --- |
| Figure 6: Free v total sulfur dioxide (red wine) | Figure 7: Free v total sulfur dioxide (white wine) |

At first glance, there appears to be a significant difference between the combined variables in the two datasets, however, the singular outlier at >250 (free) and >400 (total) in Figure 7 above proves this to be misleading. Differences do exist, however, as the white wine has consistently higher values of both variables.

|  |  |
| --- | --- |
| Figure 8: Fixed v volatile acidity (red wine) | Figure 9: Fixed v volatile acidity (white wine) |

Again, there is little difference between the two datasets when it comes to fixed v volatile acidity levels. Remembering that the white wine dataset has more than 50% more observations would account for the higher level of clustering.

|  |  |
| --- | --- |
| Figure 10: Citric acid v residual sugar (red wine) | Figure 11: Citric acid v residual sugar (white wine) |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

## Modelling

#

Table comparing CER for wine red/white with KKN where k=n

|  | Classification error rate | |
| --- | --- | --- |
| k | Red | White |
| 1 | 0.39 | 0.44 |
| 2 | 0.478 | 0.505 |
| 3 | 0.493 | 0.511 |
| 4 | 0.478 | 0.527 |
| 5 | 0.483 | 0.522 |
| 10 | 0.47 | 0.54 |
| 20 | 0.448 | 0.538 |
| 50 | 0.468 | 0.546 |
| 100 | 0.48 | 0.56 |

Therefore, based on the above (specify test set 25% and random state=4) k=1 was best scenario for both files.

Table below is classification report for both on kkn

| Classification report (red) | Classification report (white) |
| --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | *precision* | *recall* | *f1-score* | *support* | | 3 | 0.00 | 0.00 | 0.00 | 2 | | 4 | 0.10 | 0.07 | 0.08 | 14 | | 5 | 0.73 | 0.69 | 0.71 | 181 | | 6 | 0.58 | 0.62 | 0.60 | 151 | | 7 | 0.52 | 0.48 | 0.50 | 50 | | 8 | 0.12 | 0.50 | 0.20 | 2 | | **avg/total** | **0.62** | **0.61** | **0.61** | **400** | | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | *precision* | *recall* | *f1-score* | *support* | | 3 | 0.17 | 0.14 | 0.15 | 7 | | 4 | 0.20 | 0.15 | 0.17 | 41 | | 5 | 0.57 | 0.56 | 0.56 | 365 | | 6 | 0.62 | 0.62 | 0.62 | 555 | | 7 | 0.48 | 0.52 | 0.50 | 209 | | 8 | 0.47 | 0.43 | 0.44 | 47 | | 9 | 0.00 | 0.00 | 0.00 | 1 | | **avg/total** | **0.56** | **0.56** | **0.56** | **1225** | |

Decision tree

Table below is classification report for both on Decision tree – this should’ve been better, but it just wasn’t ☹

| Classification report (red) | Classification report (white) |
| --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | *precision* | *recall* | *f1-score* | *support* | | 3 | 0.00 | 0.00 | 0.00 | 2 | | 4 | 0.12 | 0.07 | 0.09 | 14 | | 5 | 0.72 | 0.65 | 0.68 | 181 | | 6 | 0.60 | 0.66 | 0.63 | 151 | | 7 | 0.56 | 0.54 | 0.55 | 50 | | 8 | 0.12 | 0.50 | 0.20 | 2 | | **avg/total** | **0.63** | **0.61** | **0.62** | **400** | | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | *precision* | *recall* | *f1-score* | *support* | | 3 | 0.00 | 0.00 | 0.00 | 7 | | 4 | 0.18 | 0.17 | 0.18 | 41 | | 5 | 0.63 | 0.63 | 0.63 | 365 | | 6 | 0.67 | 0.65 | 0.66 | 555 | | 7 | 0.53 | 0.59 | 0.56 | 209 | | 8 | 0.49 | 0.47 | 0.48 | 47 | | 9 | 0.00 | 0.00 | 0.00 | 1 | | **avg/total** | **0.60** | **0.61** | **0.61** | **1225** | |

Naïve Bayes – because this algorithm relies on the probability of each quality rating occurring equally, Naïve Bays suffered and the results were worse:

Graphs of Naïve Bayes

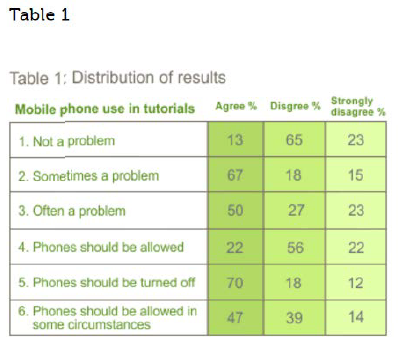
Figure 21&22

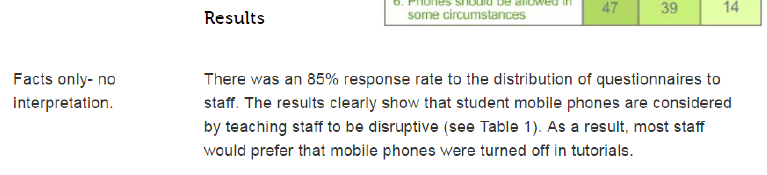
Table below is classification report for both on naïve bayes

| Classification report (red) | Classification report (white) |
| --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | *precision* | *recall* | *f1-score* | *support* | | 3 | 0.17 | 0.50 | 0.25 | 2 | | 4 | 0.07 | 0.07 | 0.07 | 14 | | 5 | 0.65 | 0.67 | 0.66 | 181 | | 6 | 0.48 | 0.38 | 0.42 | 151 | | 7 | 0.47 | 0.66 | 0.55 | 50 | | 8 | 0.25 | 0.50 | 0.33 | 2 | | **avg/total** | **0.54** | **0.54** | **0.53** | **400** | | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | *precision* | *recall* | *f1-score* | *support* | | 3 | 0.09 | 0.14 | 0.11 | 7 | | 4 | 0.25 | 0.22 | 0.23 | 41 | | 5 | 0.48 | 0.54 | 0.51 | 365 | | 6 | 0.50 | 0.33 | 0.39 | 555 | | 7 | 0.35 | 0.68 | 0.47 | 209 | | 8 | 0.50 | 0.02 | 0.04 | 47 | | 9 | 0.00 | 0.00 | 0.00 | 1 | | **avg/total** | **0.46** | **0.43** | **0.42** | **1225** | |

*Guidance:*

* *Present the empirical findings of the analysis*
* *Typically includes*
  + *Descriptive statistics*
  + *Visualisations (graphs, charts, illustrative graphics)*
  + *Analytical / model outcomes*





# Discussion

#

*Guidance:*

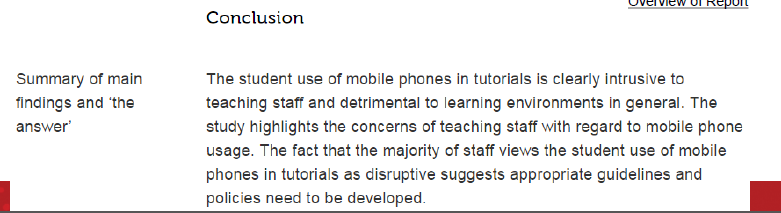
* *Presentation of main argument*
* *Explains how the results address knowledge gaps and answer the research question*

# Conclusion

#

*Guidance:*

* *Summarise findings*
* *Explain wider applicability of results*
* *Identify possible future developments and applications*
* *New research questions that have opened up*



# References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. *Modeling wine preferences by data mining from physicochemical properties.* In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[2] UCI Machine Learning Repository (2017). *Wine Quality Data Set*. Retrieved from <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>. Date retrieved: 4 May 2017.

[3] Boschetti, A and Massaron, L, 2015, *Python Data Science Essentials*, Packt Publishing Ltd, Birmingham, UK.

[4] Tutorials Point (2017). *Data Mining Decision Tree Induction.* Retrieved from <https://www.tutorialspoint.com/data_mining/dm_dti.htm>. Date retrieved: 14 May 2017.

# Appendix

Table 2: Summary statistics for red wine

|  | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 |
| mean | 8.319637273 | 0.527820513 | 0.27097561 | 2.538805503 | 0.087466542 | 15.87492183 | 46.46779237 | 0.996746679 | 3.311113196 | 0.658148843 | 10.42298311 | 5.636022514 |
| std | 1.741096318 | 0.179059704 | 0.194801137 | 1.40992806 | 0.047065302 | 10.46015697 | 32.89532448 | 0.001887334 | 0.154386465 | 0.16950698 | 1.065667582 | 0.80756944 |
| min | 4.6 | 0.12 | 0 | 0.9 | 0.012 | 1 | 6 | 0.99007 | 2.74 | 0.33 | 8.4 | 3 |
| 25% | 7.1 | 0.39 | 0.09 | 1.9 | 0.07 | 7 | 22 | 0.9956 | 3.21 | 0.55 | 9.5 | 5 |
| 50% | 7.9 | 0.52 | 0.26 | 2.2 | 0.079 | 14 | 38 | 0.99675 | 3.31 | 0.62 | 10.2 | 6 |
| 75% | 9.2 | 0.64 | 0.42 | 2.6 | 0.09 | 21 | 62 | 0.997835 | 3.4 | 0.73 | 11.1 | 6 |
| max | 15.9 | 1.58 | 1 | 15.5 | 0.611 | 72 | 289 | 1.00369 | 4.01 | 2 | 14.9 | 8 |

Table 3: Summary statistics for white wine

|  | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 4898 | 4898 | 4898 | 4898 | 4898 | 4898 | 4898 | 4898 | 4898 | 4898 | 4898 | 4898 |
| mean | 6.854787668 | 0.278241119 | 0.334191507 | 6.391414863 | 0.045772356 | 35.30808493 | 138.3606574 | 0.994027376 | 3.188266639 | 0.489846876 | 10.51426705 | 5.877909351 |
| std | 0.843868228 | 0.100794548 | 0.121019804 | 5.072057784 | 0.021847968 | 17.00713733 | 42.49806455 | 0.002990907 | 0.1510006 | 0.114125834 | 1.230620568 | 0.885638575 |
| min | 3.8 | 0.08 | 0 | 0.6 | 0.009 | 2 | 9 | 0.98711 | 2.72 | 0.22 | 8 | 3 |
| 25% | 6.3 | 0.21 | 0.27 | 1.7 | 0.036 | 23 | 108 | 0.9917225 | 3.09 | 0.41 | 9.5 | 5 |
| 50% | 6.8 | 0.26 | 0.32 | 5.2 | 0.043 | 34 | 134 | 0.99374 | 3.18 | 0.47 | 10.4 | 6 |
| 75% | 7.3 | 0.32 | 0.39 | 9.9 | 0.05 | 46 | 167 | 0.9961 | 3.28 | 0.55 | 11.4 | 6 |
| max | 14.2 | 1.1 | 1.66 | 65.8 | 0.346 | 289 | 440 | 1.03898 | 3.82 | 1.08 | 14.2 | 9 |

1. *A priori* meaning “from before” translates to say if the training data has balanced class value counts, then Naïve Bayes should work well. However, it is already known that the datasets in this study are unbalanced. [↑](#footnote-ref-1)