COSC2670: Practical Data Science

Assignment 2: Modelling wine quality based on physicochemical tests

Due midday on Thursday, 18 May 2017

Submitted by:

Casey-Ann Charlesworth (3132392)

Table of Contents

[1. An abstract/executive summary 1](#_Toc481423682)

[2. Introduction 1](#_Toc481423683)

[3. Methodology 1](#_Toc481423684)

[4. Results 2](#_Toc481423685)

[5. Discussion 4](#_Toc481423686)

[6. Conclusion 4](#_Toc481423687)

[7. References 5](#_Toc481423688)

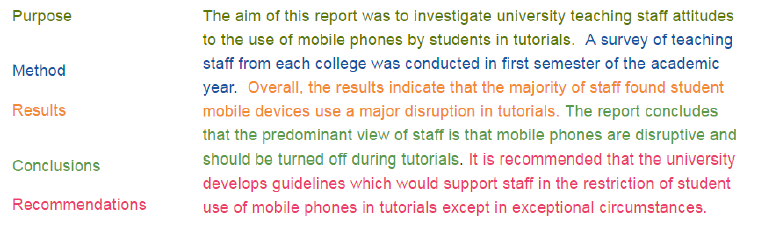
Table of Tables

Table of Figures

# An abstract/executive summary

#

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

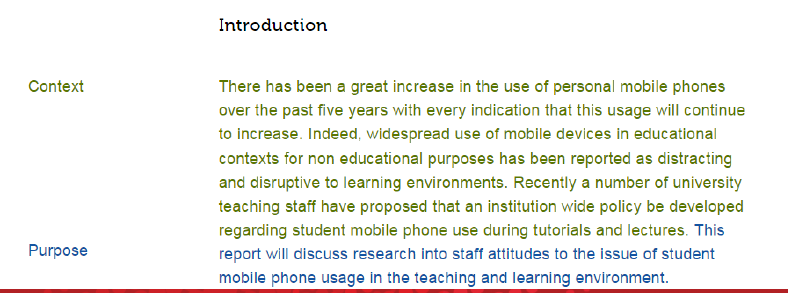


# Introduction

#

*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

The variables:

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

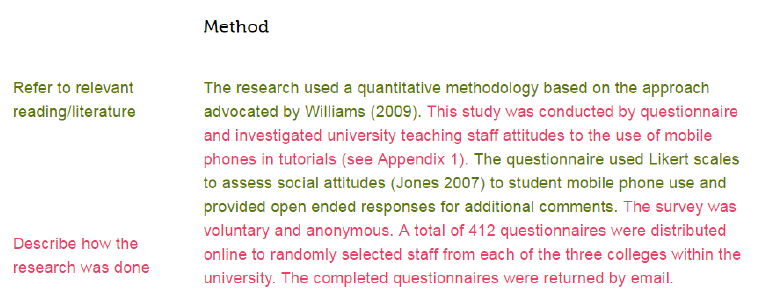
sulphates (g(potassium sulphate)/dm3)

alcohol (% vol.)

#

*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

#

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

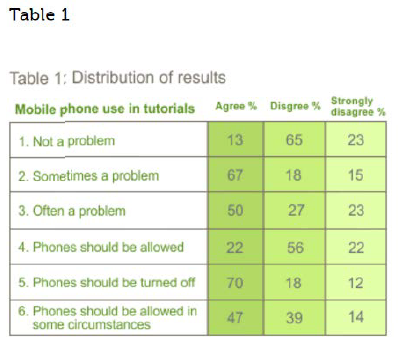
Figure 21&22

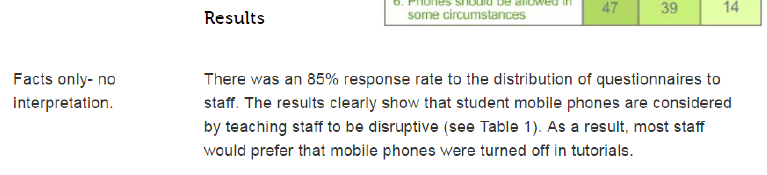
Table below is classification report for both on naïve bayes

| 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:*

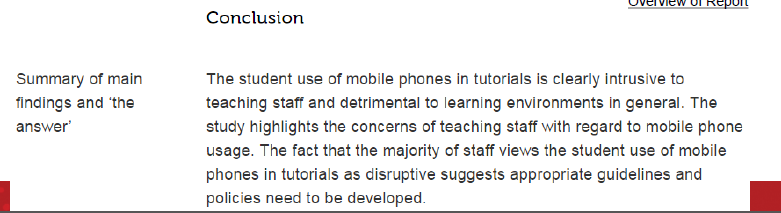
* *Presentation of main argument*
* *Explains how the results address knowledge gaps and answer the research question*

# Conclusion

#

*Guidance:*

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



# References

#

# Appendix

Table : 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 : 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 |