

**ACADEMIC**

**INTEGRITY**



Academic

Services



**ASSIGNMENT**

**DECLARATION**

**FORM**

|  |
| --- |
| **Plagiarism** means the reproduction of someone else’s words, ideas or findings and presenting them as your own ideas without proper acknowledgement and includes:  direct copying or paraphrasing from someone else’s published work (either electronic or hard copy) without acknowledging the source;  using facts, information and ideas derived from a source without acknowledgement; producing assignments (required to be independent) in collaboration with and/or using the work of other people; and assisting another person to commit an act of plagiarism. |
| **By submitting\* this piece of work and signing this document, I declare that:**  the work is my own work or the work of the group; and  I have not previously submitted all or part of this work for assessment in any subject, unless the subject coordinator for the current subject (or my research supervisor, if applicable) has given me written permission to reuse specific material and I have correctly referenced the material taken from my own earlier work; and  I have read the plagiarism avoidance advice issued to me by my Faculty; and  I have read and agree to be bound by the Statutes, Regulations and Policies of the University relating to Academic Misconduct – see [http://www.latrobe.edu.au/learning/integrity.html;](http://www.latrobe.edu.au/learning/integrity.html) and if I am discovered committing or having committed an act of academic misconduct (including plagiarism or cheating), I will be subject to student discipline processes.  I also agree that the University or any third party authorised by the University has the right to reproduce and/or communicate (make available online or electronically transmit) the work I have submitted, for the purpose of detecting work that has been copied from other sources. |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Subject Code** | C | S | E | 5 | D | M | I |  |  |
| **Subject Name** | DATA MINING | | | | |  |  |  | |
| **Lecturer’s name** | Professor Phoebe Chen | | | | |  |  |  | |
| **Name of Document** | CSE5DMI Assignment Two | | | | |  |  |  | |
| **Student Number** | 1 | 9 | 8 | 3 | 2 | 1 | 9 | 2 |  |
| **Family Name** | Palmqvist Sjövall | | | | |  |  |  | |
| **Given Names** | Anna | | | | |  |  |  | |
| **Signature of Student (Digital signature is allowed)** |  | | | | |  |  |  | |
| **Date of Signature** | 2018-10-09 | | | | |  |  |  | |

CSE5DMI Data Mining Sem 2 2018,

Assignment Two

# Part I

## Question 1a)

The original data needs preprocessing for the data to become suitable. To calculate the Accuracy and AUC we need to make the continuous values discrete, as the point of classification is to divide the data into categories unlike e.g. regression. This is done with Orange.data.preprocess.discretize.EqualWidth(n). It divides the original values into n number of bins of the same width. I chose this over EqualFreq(n) since it seems more desirable to distinct between e.g. qualities 1-3, 3-5, 5-7, 7-9 than 1-5, 5-6, 6-7, 7-9. As we have more values in the middle range of the quality spectrum, EqualFreq would give a more balanced set. However, for someone buying wine, a scale of 1-4 is preferable over being able to district between the middle qualities better. I chose n=4 as the parameter because it is the maximum value allowed (at least one class in quality has a population of 5 and the minimum number of members in a bin must be more than n).

## Question 1c)

The performance of the NN classifier is different once the minority classes are removed (<10 tuples). This removes quality classes 1,2 and 9 and allows us to increase the number of bins, n. Because the data will no longer be discretized in the same way it is not obvious to compare them to each other. If we chose n=4 in both instances, we get the split [<3] [3-5] [5-7] [>=7] for the regular data and [<4] [4-6] [6-7] [>=7] for the data without minority classes. Even though we get a lower Accuracy and AUC in the second case, this is not based on the same categories which does not render a clear conclusion. That corresponds with how the Accuracy increases when n=2, because it is an easier task to divide between ‘good’ and ‘bad’ compared to a scale 1-4.

Had another discretization method or a different number of bins been selected, which data gives the best result (highest AUC and Accuracy) differs. For example, when choosing EqualFreq instead of EqualWidth, the data without minority classes had a higher Accuracy.

A better method of testing whether minority classes negatively affects the performance of the white wine data set would have been to write an original discretizer. Most classifiers work better with balanced data, so I would still not have chosen categories 1-9. It could for example happen that the 5 values with quality 1 could all end up in the test set. Thus perhaps 6-7 bins would be appropriate where the highest and lowest ends of the classes are bundled together.

## Question 2b)

As stated in lab 6: “*Multi-layer Perceptron is sensitive to feature scaling, so it is highly recommended to scale your data”*. Therefore, the data will be scaled in the preprocessing.

## Question 2d)

Confusion matrix:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Predicted label    True label | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 1 | 2 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 28 | 9 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 230 | 136 | 3 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 206 | 321 | 14 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 34 | 184 | 5 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 4 | 40 | 2 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

We check the diagonal of the confusion matrix for the correct number of classifications (highlighted). Those are the highest for quality 5 and 6. Also, there are almost no correct classifications for the other classes. We also see there is no support for the predicted label of classes 1,2,4,8 and 9 even though they exist in the test set. Which makes sense as the support is much higher for 5,6 and 7 in the data set. We can also tell that there are many wines of quality 5 classified as 6, 6 as 5 and 7 as 6. As the diagonal values are not that much higher than the off-diagonal elements we do not have many correct predictions.

## Question 3a)

|  |  |  |
| --- | --- | --- |
|  | y=’g’ | y=’b’ |
| Cluster 0 | 157 | 33 |
| Cluster 1 | 68 | 93 |

# Part II

## Question 4a)

The following are the frequent itemsets with support greater or equal to 0.6.

|  |  |  |
| --- | --- | --- |
| Itemset | Support count | Support |
| A | 7 | 0.7 |
| C | 7 | 0.7 |
| E | 6 | 0.6 |
| F | 6 | 0.6 |
| G | 6 | 0.6 |
| H | 9 | 0.9 |
| AF | 6 | 0.6 |
| AH | 7 | 0.7 |
| CH | 7 | 0.7 |
| FH | 6 | 0.6 |
| AFH | 6 | 0.6 |

The calculations were made using Apriori’s algorithm. First, I calculated the support count for all the singleton itemsets and removed those with too low support. Then, using those, I made the possible permutations to create the itemsets with two elements. Again, removing those below the support. Finally, I checked which possible permutations were possible from {AF,AH,CH,FH} and the only one with a valid support was {AFH}.

## Question 4b)

The following are the association rules from the frequent itemsets with confidence higher or equal to 0.85.

|  |  |
| --- | --- |
| Association rule | Confidence |
| A=>F | 0.86 |
| F=>A | 1 |
| A=>H | 1 |
| C=>H | 1 |
| F=>H | 1 |
| AF=>H | 1 |
| AH=>F | 0.86 |
| FH=>A | 1 |

To calculate the confidence of X=> Y I divided the support count of XY with X. If that value was higher or equal to 0.85 it was included in the list.