

ANL303

Examination – July Semester 2019

Fundamentals of Data Mining

Tuesday, 12 November 2019

10:00 am – 12:00 pm

Time allowed: 2 hours

INSTRUCTIONS TO STUDENTS:

1. This examination contains **THREE (3)** questions and comprises **EIGHT (8)** printed pages (including cover page).
2. You must answer **ALL** questions.
3. All answers must be written in the answer book.
4. This is a **closed-book examination**.

At the end of the examination

Please ensure that you have written your examination number on each answer book used.

Failure to do so will mean that your work cannot be identified.

If you have used more than one answer book, please tie them together with the string provided.

**THE UNIVERSITY RESERVES THE RIGHT NOT TO MARK YOUR
SCRIPT IF YOU FAIL TO FOLLOW THESE INSTRUCTIONS.**

Answer all questions. (Total 100 marks)

Question 1

Better World Shopping mall (BWSM) is a shopping centre that specifically caters for the apparel needs of the urban area residents. Since last year, its revenue has been declining and many retail shops have decided to move out from the mall. The management of BWSM wants to study the amount of money that their customers would spend on shopping and divides them into groups for promotion. The management decided to use clustering to better profile their customers according to their demographics and spending power. You are now given a dataset (*BWSM.csv*) as shown in Table 1 to help BWSM to do this data mining project.

Table 1. Description of *BWSM.csv*

Attribute	Description	Labels/Values
CustomerID	Unique identifier of the customer	Unique code
Gender	Gender of the customer	“M” for Male / “F” for Female
Education	Education level of the customer	“1” for High school or below / “2” for Bachelor’s degree / “3” for Master’s degree or above
Age	Age of the customer	Integer measurement
Income	Annual Income of the customer	Dollar measurement
Household	Household type of the customer	“1” for Single / “2” for Couple / “3” for Family with children / “4” for Extended family (i.e. children and grandparents) / “5” for Others
VisitFrequency	How many times the customer visits BWSM per month	Integer measurement
AvgSpent	The average amount the customer spent in the BWSM per visit	Dollar measurement

- (a) With reference to the CRISP-DM framework, discuss how you plan to carry out this data mining project.

(24 marks)

- (b) Analyse the data based on the summary statistics given in Table 2.

Table 2. Summary statistics of the attributes

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Gender		Flag	--	--	--	--	--	2	589
Education		Nominal	1	3	--	--	--	3	589
Age		Continuous	17	46	35.092	7.803	-0.503	--	589
Income		Continuous	15799	62484	30129.543	10530.441	1.068	--	589
Household		Nominal	1	5	--	--	--	5	589
VisitFrequency		Continuous	0	35	1.553	2.605	7.451	--	589
AvgSpent		Continuous	0	880	221.375	89.570	2.531	--	589

- (i) Explain if there is a need to perform data transformation. (4 marks)
- (ii) Describe a scenario where z-score normalization is preferable to min-max normalisation. In your answer, differentiate these *two (2)* categories of data normalisation technique. (4 marks)
- (c) Assume that you have built two clustering models, Model A and Model B. The details of each model are given in Table 3. In each model, you are able to clearly describe the profile of each cluster. Based on Table 3, identify the model that you believe is better for deployment. Defend your choice by providing good reasons.

Table 3. Description of Model A and Model B

Description	Model A	Model B
Number of clusters	3	5
Number of clustering criteria	4	4
Ease of interpretation of the profile of each cluster	High	High
Average Silhouette coefficient	0.75	0.79
Size of each cluster	Cluster 1: 23% Cluster 2: 46% Cluster 3: 31%	Cluster 1: 17% Cluster 2: 25% Cluster 3: 26% Cluster 4: 20% Cluster 5: 12%

(6 marks)

Question 2

You are a data scientist in the Quality Control Department of a wine production company. You would like to understand the factors affecting the wine quality by developing a classification tree that can predict whether the wine is of “Low quality” or “High quality”. A dataset related to a particular type of white wine produced by your company was collected. The number of instances in the white wine samples are 4898.

In the dataset (*winequality-white.csv*), there are 11 attributes related to physicochemical properties of the wine and 1 attribute “Quality” indicating the quality of the wine. Table 4 shows the attributes in the *winequality-white.csv* and the range of each attribute.

Table 4. Description of *winequality-white.csv*

Attribute (unit)	Range
fixed acidity (g(tartaric acid)/dm ³)	3.8-14.2
volatile acidity (g(acetic acid)/dm ³)	0.1-1.1
citric acid (g/dm ³)	0-1.7
residual sugar (g/dm ³)	0.6-65.8
chlorides (g(sodium chloride)/dm ³)	0.01-0.35
free sulfur dioxide (mg/dm ³)	2-289
total sulfur dioxide (mg/dm ³)	9-440
density (g/cm ³)	0.987-1.039
pH	2.7-3.8
sulphates (g(potassium sulphate)/dm ³)	0.2-1.1
alcohol (vol.%)	8.0-14.2
quality	3-9

- (a) The quality of wine is initially determined by 30 wine experts in a scale that ranges from 0 (bad) to 10 (excellent) using blind wine tasting. The attribute “quality” in the dataset is the final wine quality score based on the 30 scores. Discuss whether the attribute “quality” should be the “mode”, “mean” or “median” of the scores given by the wine experts. Provide an explanation to support your answer.

(6 marks)

- (b) Based on the attribute “quality”, you decided to perform binning in the IBM SPSS Modeler to categorise the wine quality into two classes: Low Quality and High Quality. After the Binning node has been executed, there is a new attribute “quality_BIN” created with two bin values: “1” and “2”. Figure 1 shows the setting in the Binning Node. Discuss the purpose of binning and describe the meaning of the two bins in this context.

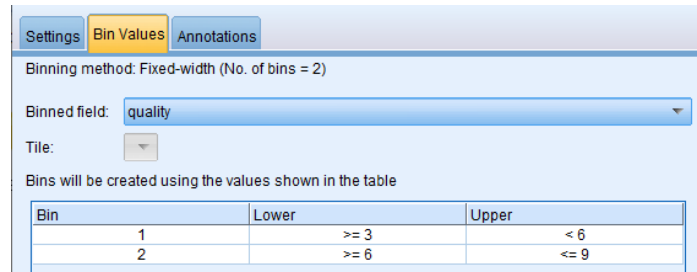


Figure 1. Setting of the Binning node

(5 marks)

- (c) You are required to analyse the dataset and build a classification tree to predict the quality of the wine based on its physiochemical properties. Before building a classification tree using the C&RT node in the IBM SPSS Modeler, the data has to be prepared by defining the measurement and role of each attribute. In your answer book, reproduce Table 5 and fill in the appropriate measurement and role for each attribute.

Table 5. Measurement and role of each attribute

Attribute	Measurement	Role
fixed acidity		
volatile acidity		
citric acid		
residual sugar		
chlorides		
free sulfur dioxide		
total sulfur dioxide		
density		
pH		
sulphates		
alcohol		
quality		
quality_BIN		

(13 marks)

- (d) Figure 2 shows a stream that you have constructed in the IBM SPSS Modeler. A Type node and a Select node have been included. Justify the inclusion of these *two* (2) nodes in the stream.

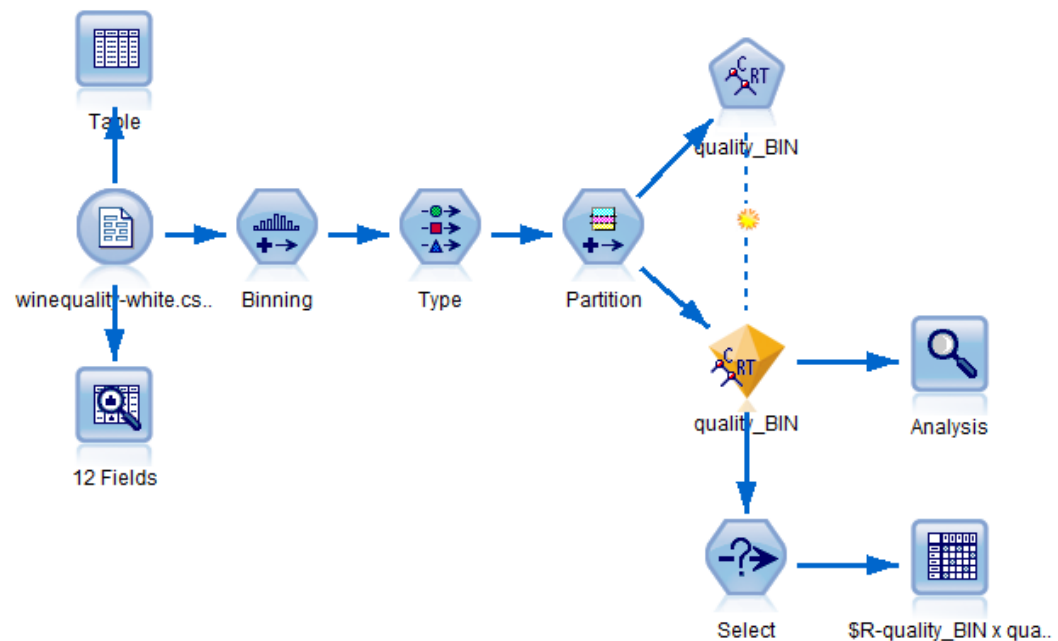


Figure 2. Stream for C&RT model and analysis

(8 marks)

- (e) In the Partition node, the training and testing partition sizes have been set to be 70% and 30%, respectively. Justify the appropriateness of this partition size setting. (6 marks)
- (f) The C&RT node has produced a decision tree as shown in Figure 3. Evaluate the quality of the decision tree model by calculating and interpreting the accuracies generated from the Matrix node as shown in Figure 4. In Figure 4, the rows show the predicted quality level and the columns show the actual quality level. Discuss whether the decision tree is useful for predicting wine quality.

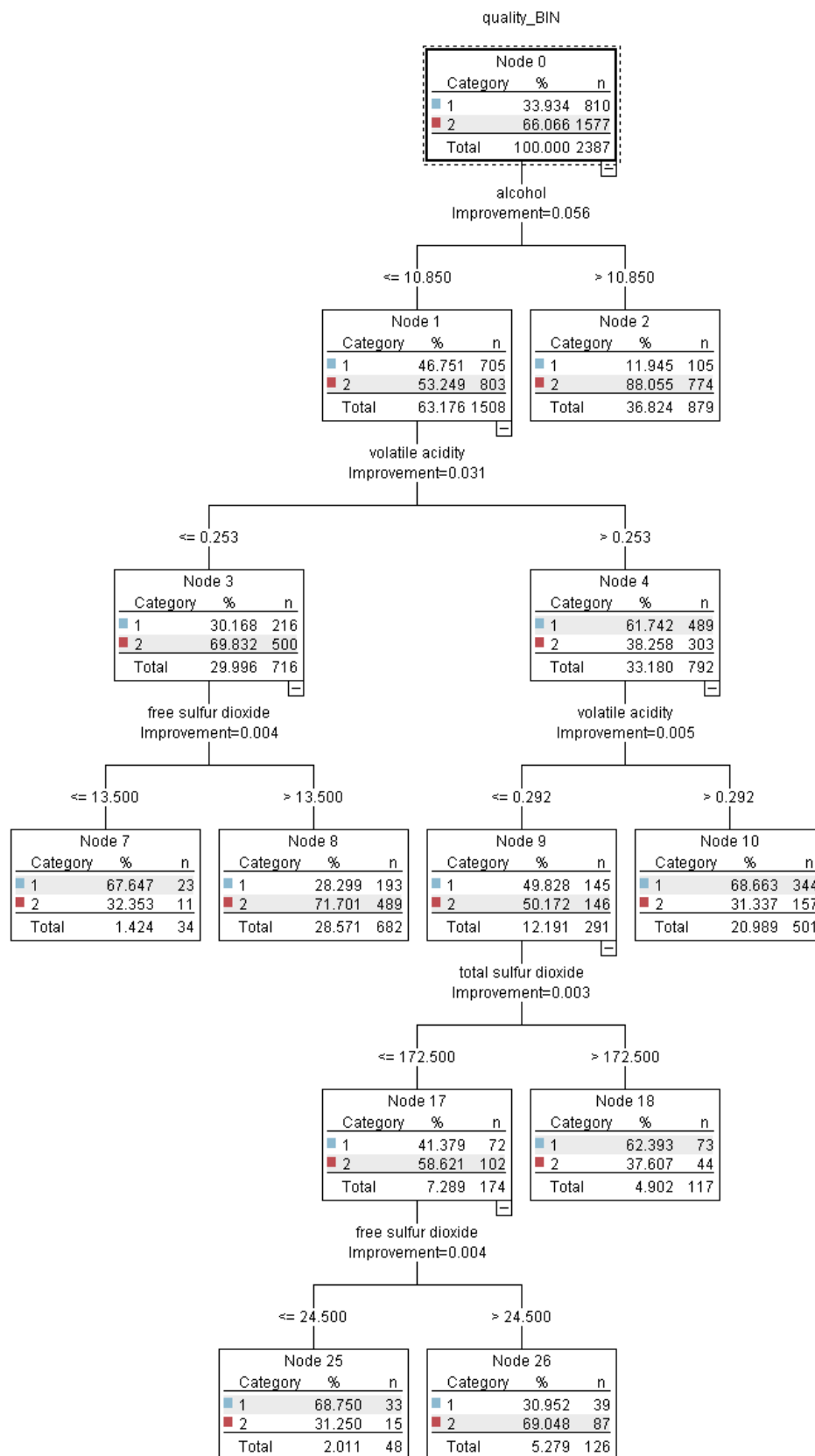


Figure 3. Decision Tree generated by the C&RT node

Training set			quality_BIN
\$R-quality_BIN	1	2	
1	653	324	
2	498	1928	

Testing set			quality_BIN
\$R-quality_BIN	1	2	
1	279	158	
2	210	848	

Figure 4. C&RT results generated from the Matrix node

(8 marks)

Question 3

Good Price Supermarket has its own warehouses where products are stored before being distributed to customers who place orders online. Currently, products within each warehouse are randomly assigned to different storage locations. Good Price Supermarket does not invest a lot in warehousing, thus all the orders are manually picked by operators. Because of the unsystematic storage assignment, operators have to walk to different sections of a warehouse to pick the products needed for fulfilling orders. As a consequence, the efficiency of the order picking process is very low.

As a data analyst, you are now asked to analyse the historical transactional data and apply data mining techniques to help the company rearrange the storage locations of products in an attempt to increase the efficiency of order picking.

- Describe an appropriate business problem where association analysis can be applied. Assess and explain why association analysis is the most suitable approach. (6 marks)
- Identify the data that the company should collect in order to conduct association analysis. (4 marks)
- Suggest **two (2)** possible limitations when applying association analysis in this scenario. (6 marks)

----- END OF PAPER -----