



UNIVERSITI TUN HUSSEIN ONN MALAYSIA

**FINAL EXAMINATION
SEMESTER I
SESSION 2019/2020**

COURSE NAME : TECHNIQUES OF DATA MINING
COURSE CODE : BWB 44603
PROGRAMME CODE : BWQ
EXAMINATION DATE : DECEMBER 2019 / JANUARY 2020
DURATION : 3 HOURS
INSTRUCTION : ANSWER ALL QUESTIONS

TERBUKA

THIS QUESTION PAPER CONSISTS OF SIX (6) PAGES

- Q1** (a) Data mining generally divided into two major categories. Define these categories and explain clearly. Give at least **TWO (2)** examples for each category. (8 marks)
- (b) Briefly comments why the researchers more interested in 90% of confidence instead of 100% confidence in association rules. (3 marks)
- (c) Differentiate between nominal, ordinal, interval and ratio attributes. Give **ONE (1)** example for each attribute. (8 marks)
- (d) Discuss whether the following activities is a data mining task. Justify your answer. (12 marks)
- (i) Grouping the students according to their CGPA.
 - (ii) Forecast the total sales of a company.
 - (iii) Monitoring the conditions of the patients in ICU.
 - (iv) Classified the level of injury in an accident.
 - (v) Predicting the weather for a week.
 - (vi) Sorting the matrix cards number based on the alphabet.

- Q2** (a) Discuss the use of *Support* and *Confidence* in association rules. (4 marks)
- (b) **Table Q2(b)** shows the market basket transactions in Supermarket Value. Calculate the confidence for the association rules $\{a,b\} \rightarrow \{e\}$, $\{e\} \rightarrow \{a,b\}$ and $\{a\} \rightarrow \{b,e\}$. Justify your results.

Table Q2(b)

Customer ID	Transaction ID	Items Bought
1	0001	{a, b, d, e, f, g}
2	0024	{a, b, c, d, e, g}
3	0012	{a, b, d, f, g}
4	0031	{a, c, d, f, g}
5	0015	{a, b, c, d, e}
6	0022	{b, d, f}
7	0029	{c, d, g}
8	0040	{a, b, c, d, e}
9	0033	{a, b, e, f, g}
10	0038	{a, b, e, f}

- (9 marks)
- (c) From **Figure Q2(c)**, summarizes **FIVE (5)** of the market basket analysis. (5 marks)

TERBUKA

Q3 (a) **Figure Q3(a)** shows the output of decision tree from a credit card data set. The class label is defined as whether the respondent is a student (yes) or not student (no). Answers the following questions.

(i) Define the most contributed factors of applying the credit card. Explain clearly your answer. (2 marks)

(ii) Select **FIVE (5)** branches and elaborate these branches. (10 marks)

(iii) Conclude by giving **THREE (3)** point from this decision tree. (6 marks)

(b) **Table Q3(b)** summarizes a data set with three attributes A, B and C with two labels (**C1** and **C2**). Construct a two-level decision tree by using the classification error rate. Show the contingency table and information gains for each of the error rate.

Table Q3(b)

A	B	C	Number of Instances	
			C1	C2
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

(13 marks)

Q4 (a) K-means is one of the several data mining techniques. Discuss **TWO (2)** issues or limitations of this technique. (4 marks)

(b) **Figure Q4(b)** shows the output of k-means clustering from iris data set. Summarize by giving **THREE (3)** points from this figure. (6 marks)

(c) Perform complete and average linkage of hierarchical clustering by using the similarity matrix in **Figure Q4(c)**. Show your results by constructing a clear dendrogram and nested clustering. (10 marks)

– END OF QUESTIONS –

TERBUKA

FINAL EXAMINATION

SEMESTER / SESSION : SEM I / 2019/2020
 COURSE NAME : TECHNIQUES OF DATA MINING

PROGRAMME CODE : BWQ
 COURSE CODE : BWB 44603

lhs	rhs	support	confidence	lift	count
[1] {WOBBLY CHICKEN}	=> {METAL}	0.001261773	1	443.82000	28
[2] {WOBBLY CHICKEN}	=> {DECORATION}	0.001261773	1	443.82000	28
[3] {DECOURPAGE}	=> {GREETING CARD}	0.001036456	1	389.31579	23
[4] {BILLBOARD FONTS DESIGN}	=> {WRAP}	0.001306836	1	715.83871	29
[5] {WOBBLY RABBIT}	=> {METAL}	0.001532153	1	443.82000	34
[6] {WOBBLY RABBIT}	=> {DECORATION}	0.001532153	1	443.82000	34
[7] {FUNK MONKEY}	=> {ART LIGHTS}	0.001712406	1	583.97368	38
[8] {ART LIGHTS}	=> {FUNK MONKEY}	0.001712406	1	583.97368	38
[9] {BLACK TEA}	=> {SUGAR JARS}	0.002072912	1	238.61290	46
[10]{BLACK TEA}	=> {COFFEE}	0.002072912	1	69.34687	46

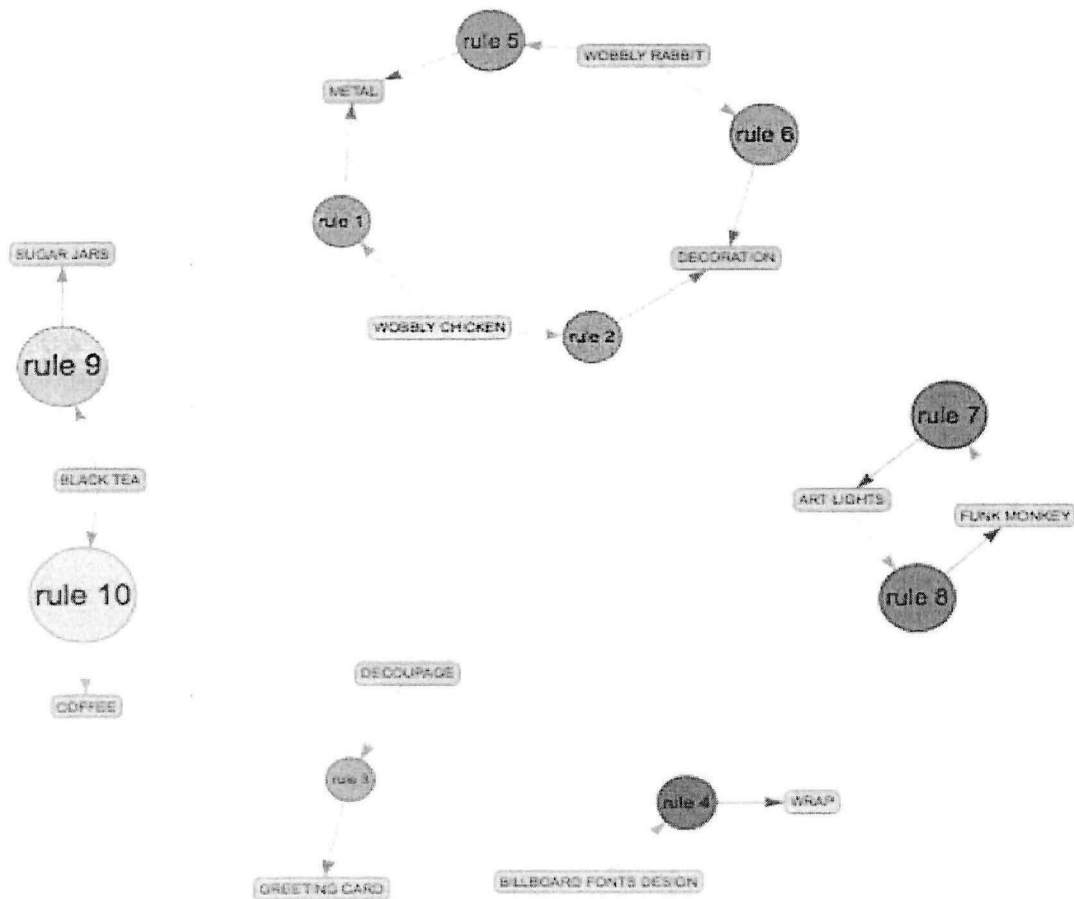


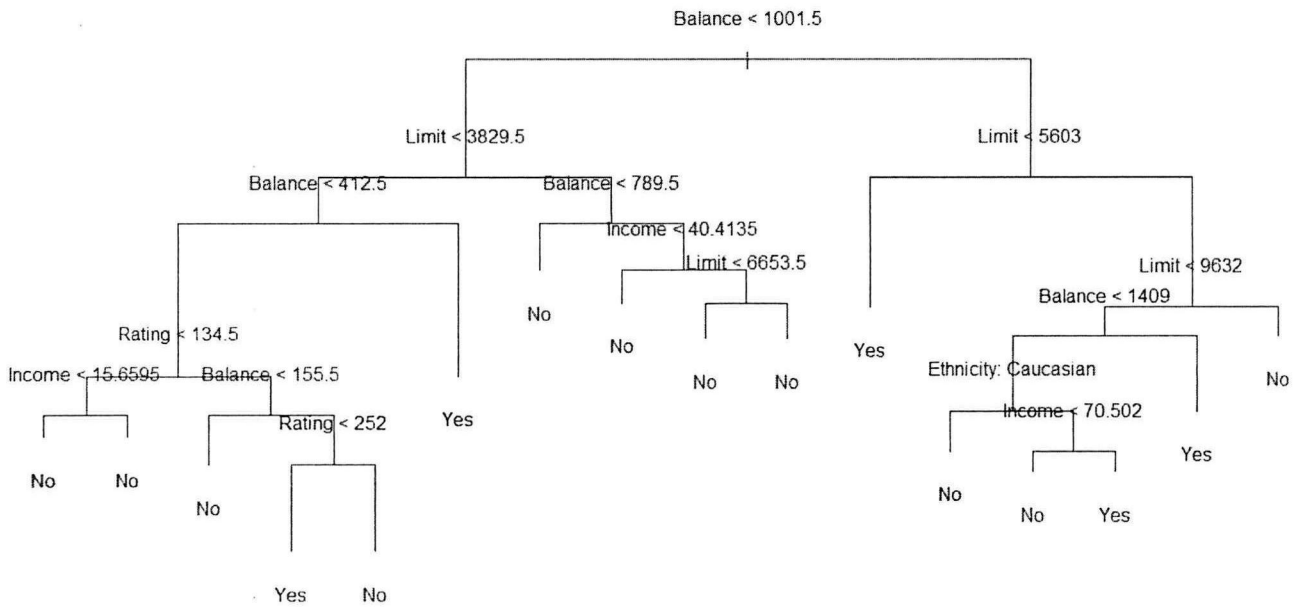
Figure Q2(c)

TERBUKA

FINAL EXAMINATION

SEMESTER / SESSION : SEM I / 2019/2020
 COURSE NAME : TECHNIQUES OF DATA MINING

PROGRAMME CODE : BWQ
 COURSE CODE : BWB 44603



```

> names(credit)
[1] "Income" "Limit" "Rating" "Cards" "Age" "Education" "Gender"
[8] "student" "Married" "Ethnicity" "Balance"
    
```

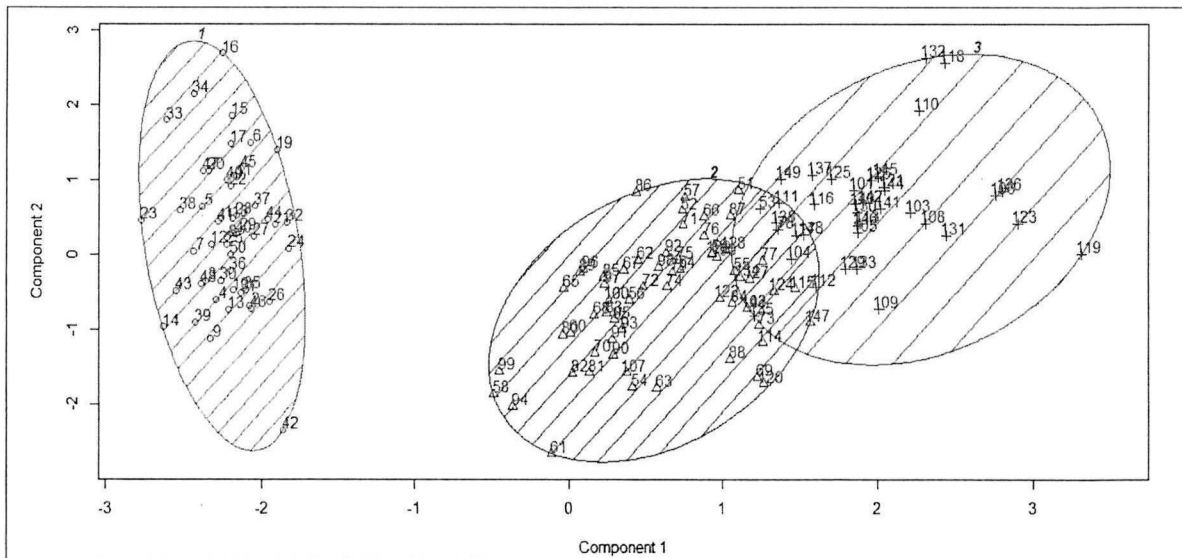
Figure Q3(c)

TERBUKA

FINAL EXAMINATION

SEMESTER / SESSION : SEM I / 2019/2020
 COURSE NAME : TECHNIQUES OF DATA MINING

PROGRAMME CODE : BWQ
 COURSE CODE : BWB 44603



These two components explain 95.81 % of the point variability.

```
table(iris$Species, km_iris$cluster)
```

	1	2	3
setosa	50	0	0
versicolor	0	48	2
virginica	0	14	36

Figure Q4(b)

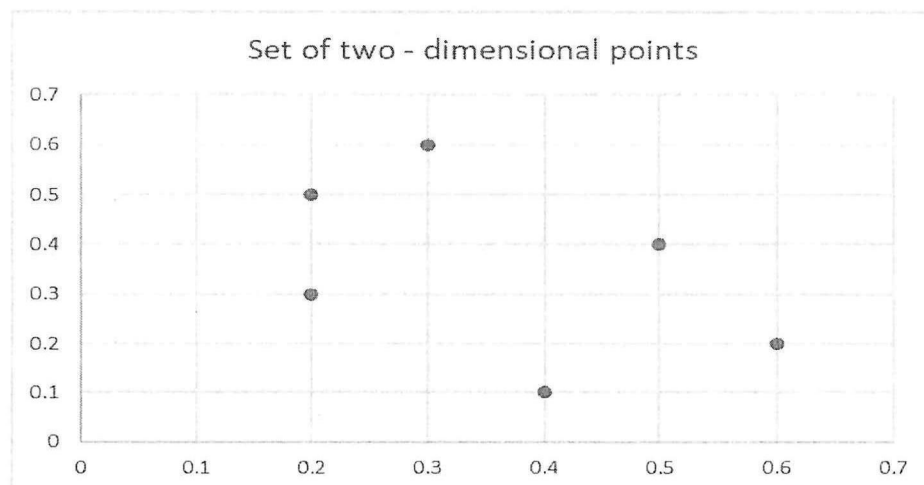


Figure Q4(c)

TERBUKA