Charles Darwin University

Final Examination

Family Name						
Given Name/s						
Student Number						
Teaching Period	Semester 1, 2019					

PRT564 – Data Analytics and Visualisation	DURATION	
	Reading Time:	10 minutes
	Writing Time:	180 minutes

INSTRUCTIONS TO CANDIDATES

- The examination has **SIX** questions. **ALL** questions must be answered on the Answer Booklet provided.
- Please ensure that your name and student number are clearly indicated **both** on your Answer Sheet and at the top of this examination paper.
- The total marks of this examination are **100 marks**.

EXAM CONDITIONS

You may begin writing from the commencement of the examination session. The reading time indicated above is provided as a guide only.

This is a RESTRICTED OPEN BOOK examination

Any non-programmable calculator is permitted

One A4 sheet of handwritten single-sided notes permitted

Hard copy, unannotated English translation dictionary only

ADDITIONAL AUTHORISED MATERIALS	EXAMINATION MATERIALS TO BE SUPPLIED
No additional printed material is permitted	1 x 20 Page Book 1 x Scrap Paper

THIS EXAMINATION IS PRINTED DOUBLE-SIDED.

THIS PAGE HAS BEEN INTENTIONALLY LEFT BLANK.

Question 1: Data Analytics and Visualisation Concepts (6 + 5 + 4 = 15 marks)

Q1a.

Choose and answer ONE of the questions below:

1. Briefly discuss the three characteristics of Big Data. Provide examples if necessary.
2. Briefly discuss the key skill sets and the profile of a data scientist.

(Marks: 6)

Q1b.

Briefly explain the key difference between unsupervised and supervised machine learning (ML). Justify with reasons which type or types of ML does deep learning belong to?

(Marks: 5)

Q1c.

Choose and answer ONE of the questions below:

1. Give two examples of diagrams to create in the final presentation of an analytics project and briefly explain to which audience each diagram is appropriate.
2. Name three typical diagrams for data visualization. Identify which diagram would be appropriate to show data changing over time.

(Marks: 4)

Question 2: Basic Data Analytics and Visualisation in R (6 + 4 + 4 + 5 + 4 + 2 = 25 marks)

Q2a.

For each sets of typed commands below, write down the output (if any) in the R console window:

1.

```
> x<-c(1,1,2,3,5,8,13)
> x[x>4]
```

2.

```
> matrix(data=c(9,2,3,4,5,6),ncol=3)
```

3.

```
> h = seq(from=1, to=5)
> for(i in 2:6)
  {
        print(h[i] * 10)
  }
```

(Marks: 6)

Q2b.

What is the R function used to encode a vector as a category? What does the function rnorm(100) do?

(Marks: 4)

Q2c. Write R codes to generate a data frame d as shown below and then calculate the mean value of its column z.

	x	y	z
1	11	19	10
2	12	20	9
3	14	21	7

(Marks: 4)

Q2d.

Briefly explain why Hexbinplot is better than Scatterplot for big data. In addition, explain what a Scatterplot Matrix is and what R function is used to generate Scatterplot Matrix.

(Marks: 5)

Q2e.

Choose and answer ONE of the questions below:

1. Briefly and separately explain how hypothesis testing can be used for model evaluation as well as model building and planning.
2. Briefly explain what t-test/t-statistic is and how it differs from the Wilcoxon Rank Sum Test.
3. Briefly explain what ANOVA-test is and what the F statistics is.

(Marks: 4)

Q2f.

You are analysing two normally distributed populations, and your null hypothesis is that the mean m1 of the first population is equal to the mean m2 of the second. Assume the significance level is set at 0.05. If the observed p value is 9.33e-02, what will be your decision regarding the null hypothesis?

(Marks: 2)

Question 3: Clustering and Association Rule Mining (8 + 8 = 16 marks)

Q3a.

Given a set of one-dimensional points: {1, 1, 2, 3, 5, 8, 13, 21} and two random initial centroids 0 and 11, perform two iterations of k-means algorithm on these points. Show your work in each iteration by writing down i) the assignment of points to centroids; ii) the updated centroid points.

(Marks: 8)

Q3b.

Given the following transaction table and the minimum support of 0.5, apply the Apriori Algorithm for two iterations to list out all the frequent itemsets of sizes 1 and 2. Show your work in each iteration by writing down both the candidate itemsets and the itemsets after pruning.

TID	Items
1	I1, I3, I4
2	I2, I3, I5
3	I1, I2, I3, I5
4	I2, I5

(Marks: 8)

Question 4: Linear Regression and Logistic Regression (4 + 5 + 5 = 14 marks)

Q4a.

What is the R function used for linear regression? And what is the R function used for logistic regression? In addition, if b = -0.5 is an estimated coefficient of a variable x in a logistic regression model, what is the effect on the odds ratio for every one unit increase in the value of x?

(Marks: 4)

Q4b.

Given the following confusion matrix from a classifier, compute Accuracy, False Positive Rate, False Negative Rate, Precision and Recall.

	Predicted Classes		
Actual Classes	Positive	Negative	Total
Positive	262	38	300
Negative	29	671	700
Total	291	709	1000

(Marks: 5)

Q4c.

Briefly explain i) how N-Fold cross validation method is used for diagnosing a fitted model and ii) how ROC curve is used to diagnose the effectiveness of a logistic regression model?

(Marks: 5)

Question 5: Naïve Bayes and Decision Tree (9 + 8 = 17 marks)

Q5a.

The following table provides the details of all the previous patients attended by the doctor (the Training Data). The table includes binary and multi-category symptoms with abbreviated names and in the last column the corresponding flu diagnoses.

	Chills (C)	Runny nose (RN)	Headache (H)	Fever (F)	Flu
Training Data	1	0	Mild	1	0
	1	1	No	0	1
	1	0	Strong	1	1
	0	1	Mild	1	1
	0	0	No	0	0
	0	1	Strong	1	1
	0	1	Strong	0	0
	1	1	Mild	1	1

Test Data	1	1	Strong	0	?

Using Naïve Bayes Classifier to predict the Test Data above for the doctor, that is, decide if a new patient with her symptoms has a flu. Show your work to get partial marks even if the calculations contain error. You can only get 1 mark for a correct guess without showing work.

(Marks: 9)

Q5b.

For building a decision tree from the training data in Q5a, select the root to split on between attributes Chills (C) and Fever (F) (assuming we ignore the others) by calculating their information gains. Show your work to get partial marks even if the calculations contain error.

(Marks: 8)

Question 6: Time Series Analysis and Text Analysis (7 + 6 = 13 marks)

Q6a.

Briefly explain what the key components of a time series are and also what the key components are in the ARIMA model.

(Marks: 7)

Q6b.

The words 'car', 'auto, 'best' have the following frequencies in ten documents Doc1, …, Doc10. Compute the tf-idf scores for each of these terms in Doc9 and then conclude which term is the most important to Doc9. Simply assume that all other words in the documents are stop words.

Term\Doc	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
car	3	0	0	5	12	0	0	2	5	1
auto	8	6	4	12	0	0	9	1	8	10
best	0	1	7	0	1	5	12	0	2	0

(Marks: 6)

END OF THE EXAMINATION.

THIS PAGE HAS BEEN INTENTIONALLY LEFT BLANK.