Name: _____ Admission No: _____

Class: _____

SINGAPORE POLYTECHNIC
2019/2020 SEMESTER ONE END SEMESTER TEST

MS2215/MS4215/MS6215

STATISTICS AND ANALYTICS FOR ENGINEERS

Time Allowed: 1 hour 30 min + 10 min reading time

Instructions to Candidates:

a) The Singapore Polytechnic examination rules are to be complied with. Any candidate who cheats or attempts to cheat will face disciplinary action and is liable to be expelled from the Polytechnic.
b) This paper consists of 15 printed pages including the cover page.
c) Answer **ALL** questions on this question paper itself. This paper consists of two sections. Total marks for Section A is 50. Total marks for Section B is 50.
d) The total marks for this paper is 100.
e) Give all non-exact answers to 3 significant figures.
f) Do not turn over this cover sheet until you are told to do so.
g) This test requires a data file which can be downloaded from **Blackboard** > MS_SAE > Learning Resources > CA Data folder. Please follow your invigilator's instructions.
h) You are allowed one double-sided A4-sized handwritten notes for reference and the use of a laptop with the required software installed (Excel, Minitab Express, KNIME). Sharing is not allowed.

Section	Question	Marks
A	1	/10
	2	/10
	3	/15
	4	/15
B	5	/10
	6	/24
	7	/16
	Total	/100

Section A (50 marks)

There are 4 questions in this Section A that require the use of software.
For Questions 1, 2 and 3, use the data file "seeds_dataset.csv".
For Question 4, no data file is provided.
Below is the description of the "seeds_dataset.csv".

The seeds dataset (Resource: UCI Machine Learning Repository) contains the measurements of geometrical properties of kernels belonging to three different species of seeds (Group).

The meaning of the features are given as follows:

Feature	**Meaning**
Area, A	Area of seed image
Perimeter, P	Perimeter of the seed image
Compactness	Derived quantity given by $4\pi A/P^2$
Length	Length of kernel
Width	Width of kernel
Asymmetry coefficient	A measure of how asymmetrical the image of the seed is
Length of kernel groove	Length of kernel groove
Group	Species of seed: 'Kama', 'Rosa' and 'Canadian'

Question 1 (10 marks)
The data in the Excel file "seeds_dataset.csv" will be used to build a model for the classification of seeds into 'Kama', 'Rosa' or 'Canadian' group.

a) How many records are there in the dataset? (1 mark)

b) Is the task described above a data query task or a data mining task? (1 mark)

c) Using KNIME or otherwise, determine the following quantities. (8 marks)

(i) Overall mean of the **Area** feature:

(ii) Mean area for each group of seed

Canadian group mean area:

Kama group mean area:

Rosa group mean area:

(iii) Construct a **box-and-whisker plot** showing the distribution of **Area** for each **Group**. Based on your plot, comment on whether **Area** alone can be used to classify the different seeds into their respective species. (*Note: You do not have to draw the chart here.*)

(iv) Construct a **Scatterplot** between the **Area** and the **Length of kernel groove** using a colour scheme to distinguish between different **Group** (seed species) in the plot. Do you think these 2 features are significant in distinguishing between the species 'Rosa' and 'Canadian'? Explain. (*Note: You do not have to draw the chart here.*)

Question 2 (10 marks)

Perform K-Means clustering on the seeds dataset to uncover the clusters for the species. In the Partitioning node, choose the size of first partition as relative 99% drawn randomly using the random seed of '2019'. The first partitioned data is to be normalized using min-max (0.0-1.0) and used to create the K-means model. The second partitioned data will be used for testing.

a) Explain why the value of *k* to be used is 3. (1 mark)

b) Explain why the clustering is done on normalized data. (2 marks)

c) The number of records in each of the resulting 3 clusters. (1.5 marks)

d) The resulting cluster centroids for each cluster (Fill in the blanks). (2.5 marks)

Row ID	Area	Perimeter	Compactness	Length of kernel	Width of kernel	Asymmetry coefficient	Length of kernel groove
cluster_0	0.76		0.697		0.773	0.365	0.759
cluster_1	0.383	0.419		0.363		0.261	0.317
cluster_2	0.123	0.174	0.377	0.186	0.161		0.279

e) Based on the model created in (a), indicate the cluster assigned to the following data in the second partition below using the cluster ID used in (b): (3 marks)

Row ID	Area	Perimeter	Compactness	Length of kernel	Width of kernel	Asymmetry coefficient	Length of kernel groove	Cluster ID
Row 10	15.26	14.85	0.87	5.714	3.242	4.543	5.314	
Row 75	16.77	15.62	0.864	5.927	3.438	4.92	5.795	
Row 163	12.55	13.57	0.856	5.333	2.968	4.419	5.176	

Question 3 (15 marks)

Build a classification model to predict the species of seeds using the decision tree method with 'Gini index' (no pruning). Reduced error pruning is checked to cut the tree in post-processing by the algorithm in KNIME. In the Partitioning node, choose the size of first partition as relative 80% drawn randomly using the random seed of '2019'. Restrict the minimum number of records per node to be 10.

a) Which is the target variable? (1 mark)

b) The workflow in *Figure 1* is created to perform the classification task. Name each of the nodes below. (5 marks)

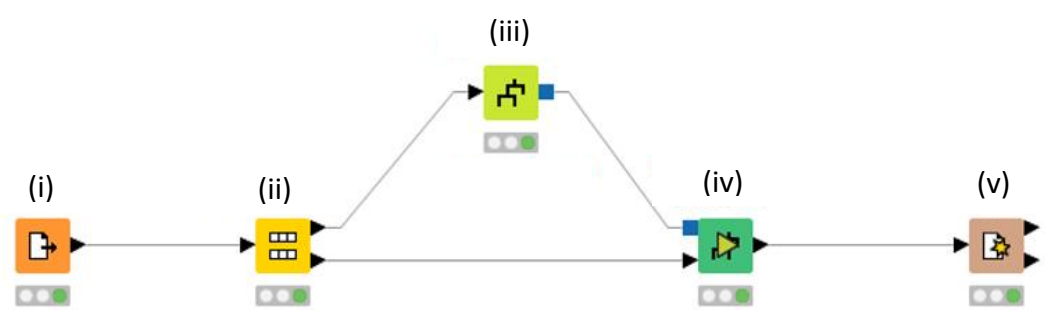


Figure 1

Node (i):

Node (ii):

Node (iii):

Node (iv):

Node (v):

c) From the confusion matrix generated by the workflow in *Figure 1*, how many False Negative for class 'Kama' are produced? Interpret your answer in context. (3 marks)

d) Based on the decision tree produced above, summarize the tree as a set of classification rules of the form "IF… THEN…" for each species. Each rule should include a **probability greater than 0.90**. (6 marks)

IF THEN

IF THEN

IF THEN

Question 4 (15 marks)

Blood pressure readings have two numbers. The first number is the systolic blood pressure. The second number is the diastolic blood pressure. For example, 120/80 mmHg reads a systolic blood pressure of 120 mmHg and diastolic blood pressure of 80 mmHg. A doctor took blood pressure measurements (in mmHg) of 14 patients (refer to *Table 1*) and he obtained the linear regression results with the values masked out (refer to the *Table 2*):

Systolic	Diastolic
138	82
130	91
135	100
140	100
120	80
125	90
120	80
130	80
130	80
144	98
143	105
140	85
130	70
150	100

Table 1

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Systolic				
Intercept				

Multiple R-Squared: 0.4328
Adjusted R-Squared: 0.3855

Table 2

a) Identify the predictor and response variables based on *Table 2*. (2 marks)

b) Build a simple linear regression model using KNIME. Write down the equation of the model. (5 marks)

c) Interpret the equation of the model obtained in (b). State also if the 'Intercept' value has any meaning in this context. (3 marks)

d) Estimate the diastolic pressure given that the systolic pressure is 120 mmHg. (3 marks)

e) Is the equation useful to estimate diastolic pressure? Explain in context. (2 marks)

Section B: (50 marks)

This section does not require the use of any software. Answer all questions in the space provided. Show all your workings clearly.

Question 5 (10 marks)

Given a set { 2, 3, 4, 10, 11, 12, 20, 25, 30 } is clustered using K-Means with k = 2.

The centroids of the two clusters are initially chosen as Centroid_1 = 4 and Centroid_2 = 12 respectively.

a) For the first round: (4 marks)

(i) Euclidean distance between '10' and Centroid_1:

(ii) Euclidean distance between '10' and Centroid_2:

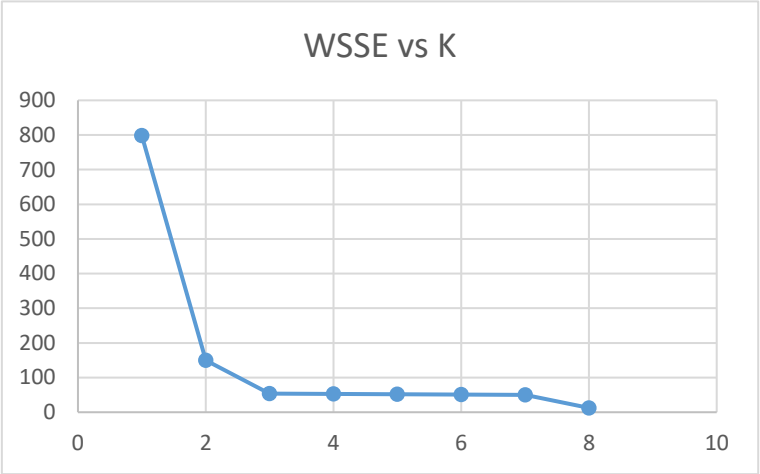
(iii) Hence, state the cluster for which '10' is assigned to and explain why.

b) After the first round, numbers '2' and '3' are assigned to Cluster 1, while numbers '10', '11', '20', '25' and '30' are assigned to Cluster 2. (4 marks)

(i) Calculate the new centroid of Cluster 1. (Show your workings clearly.)

(ii) Calculate the new centroid of Cluster 2. (Show your workings clearly.)

c) The following is a graph of within-cluster sum of squared error (*WSSE*) vs *K* in the K-Means clustering process, which *K* is the best number of cluster? Explain in context. (2 marks)



Question 6 (24 marks)
The data below is used to predict the annual income (<=50K or >50K) of individuals, based on the attributes *Education* (years) and *Marital_Status* (Others or Married).

Education	Marital_Status	Income
3	Married	<=50
6	Married	<=50
9	Married	<=50
9	Married	<=50
9	Married	<=50
9	Others	<=50
9	Married	<=50
9	Others	>50
10	Married	<=50
10	Others	<=50
10	Others	<=50
10	Others	<=50
10	Others	<=50
13	Married	>50
13	Married	>50
13	Married	<=50
13	Others	>50
13	Others	<=50
13	Others	<=50
13	Others	<=50

a) Use the data above to estimate each probability below. (10 marks)

$P(<=50K) =$ $P(>50K) =$

$P(Education \le 9yrs \mid <=50K) =$ $P(Education \le 9yrs \mid >50K) =$

$P(Education > 9yrs \mid <=50K) =$ $P(Education > 9yrs \mid >50K) =$

$P(Marital\_Status = Married \mid <=50K) =$ $P(Marital\_Status = Married \mid >50K) =$

$P(Marital\_Status = Others \mid <=50K) =$ $P(Marital\_Status = Others \mid >50K) =$

b) Use the naïve Bayes approach to predict the annual income (<=50K or >50K) of someone with more than 9 years of education and is *not* married. (12 marks)

c) Given an individual who earns more than 50K a year, what is the probability that he/she has more than 9 years of education and is not married, using naïve Bayes assumption of independence? (2 marks)

Question 7 (16 marks)

In the statistical modelling of 'Sales' (Y) of a certain electronic product, 6 predictor variables were used. They are, 'Time to Market' (X1), 'Market Potential' (X2), 'Advertisement Expenditure' (X3), 'Market Share' (X4), 'Change Introduced by Product' (X5) and 'Design Complexity' (X6).

In the first stage of modelling the 'Sales', only X1, X2, X3 and X4 were used. The results of the modelling process and the relevant statistics are given below.

Variable	Coeff.	Std. Err.	t-value	P>|t|
Time to Market (X1)	3.8165	1.2698	3.0055	0.007
Market Potential (X2)	0.0444	0.0072	6.2024	4.66E-6
Advertisement Expenditure (X3)	0.1525	0.038	4.0144	0.0007
Market Share (X4)	259.4698	42.1826	6.1511	5.21E-6
Intercept	-1,312.2922	440.7472	-2.9774	0.0074

Multiple R-Squared: 0.896
Adjusted R-Squared: 0.8752

Table 3

a) Based on *Table 3*, answer the following:

(i) Construct a linear model to estimate 'Sales' (Y) in terms of X1, X2, X3 and X4. (3 marks)

(ii) Identify the most influential variable. Explain your choice. (2 marks)

(iii) Interpret the numerical meaning of the coefficient of X4. (2 marks)

(iv) Estimate the 'Sales' for the following combination of the predictor variable values: X1 = 40 units, X2 = 50000 units, X3 = 6000 units, X4 = 12 units. (3 marks)

(v) State the R^2 value and interpret the value in context. (2 marks)

b) Another model was built for 'Sales' (Y), with variable X5 'Change Introduced by Product', included as a predictor variable. The relevant results of the modelling process are given below.

Variable	Coeff.	Std. Err.	t-value	P>|t|
Time to Market (X1)	3.6121	1.1817	3.0567	0.0065
Market Potential (X2)	0.0421	0.0067	6.2527	5.27E-6
Advertisement Expenditure (X3)	0.1289	0.037	3.4792	0.0025
Market Share (X4)	256.9556	39.1361	6.5657	2.76E-6
Change Introduced by Product(X5)	324.5335	157.2831	2.0634	0.053
Intercept	-1,113.788	419.8869	-2.6526	0.0157

Multiple R-Squared: 0.915
Adjusted R-Squared: 0.8926

(i) Explain what is described by an adjusted R^2 value. (2 marks)

(ii) Comment on the impact of the contribution of X5, to the predictive strength of the model. (2 marks)

***** END OF PAPER *****