**SINGAPORE POLYTECHNIC**
**2020/2021 SEMESTER TWO END SEMESTER TEST**
**MS2215/MS4215/MS6215**

# STATISTICS AND ANALYTICS FOR ENGINEERS

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

| Name: | |
|---|---|
| Class: | Admission No: |

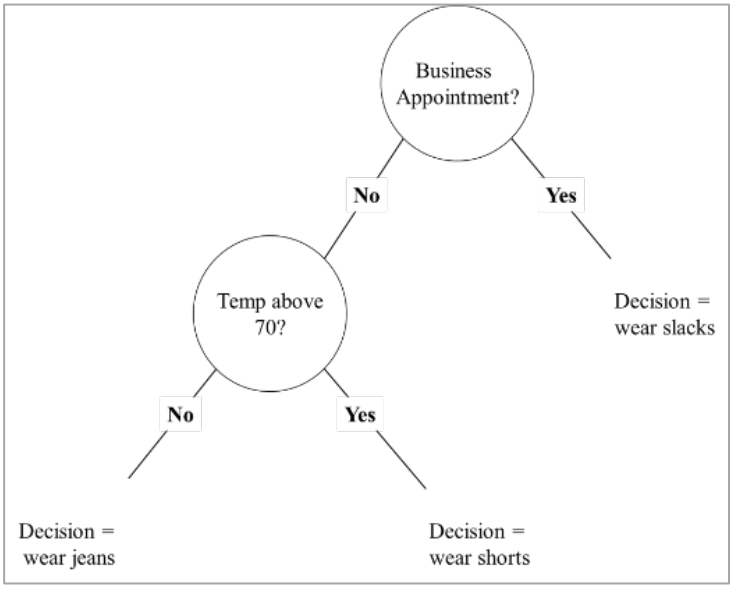
Instructions to Candidates:

1. 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.
2. This paper consists of **12** printed pages including the cover page. Answer **ALL** questions on this question paper itself.
3. Give all non-exact answers to **3 significant figures**. The total marks for this paper is 100.
4. Do not turn over this cover sheet until you are told to do so.
5. This test requires two data files (for Question 1 and Question 3) which can be downloaded from **Blackboard** > MS_SAE > Learning Resources >CA Data folder. Please follow your invigilator's instructions.
6. You are allowed one double-sided A4-sized handwritten notes for reference and laptop with Minitab Express, Microsoft Excel and KNIME installed. Sharing is not allowed.

| Question | Marks |
|---|---|
| MCQ | /10 |
| 1 | /25 |
| 2 | /20 |
| 3 | /25 |
| 4 | /20 |
| Total | /100 |

**Multiple Choice Questions (10 marks)**

<table>
<tr><td>1</td><td>Supervised learning differs from unsupervised clustering in that supervised learning requires _____.<br>a. at least one input attribute<br>b. input attributes to be categorical<br>c. at least one output attribute<br>d. output attributes to be categorical</td><td>( )</td></tr>
<tr><td>2</td><td>Which <u>one</u> of the following is the <u>correct</u> order described in the CRISP-DM (Cross-Industry Standard Process for Data Mining) process?
<table>
<tr><td>i</td><td>Evaluation</td><td>ii</td><td>Business Understanding</td><td>iii</td><td>Data Preparation</td></tr>
<tr><td>iv</td><td>Data Understanding</td><td>v</td><td>Deployment</td><td>vi</td><td>Modelling</td></tr>
</table>
a. iv, iii, ii, vi, i and v<br>b. ii, iv, iii, vi, i and v<br>c. iv, ii, iii, vi, i and v<br>d. ii, iii, iv, v, vi and i</td><td>( )</td></tr>
<tr><td>3</td><td>Marketing is crucial for the growth and sustainability of any business. However, one of the key pain points for any marketing professionals is to know the customers and identify their needs. Tasked by colleagues in the marketing department, your job is to create a model to perform customer grouping based on the underlying demographics data of the customer. What could be the most appropriate technique adopted?<br>a. Regression<br>b. Clustering<br>c. Classification<br>d. Data query</td><td>( )</td></tr>
</table>

| | | |
|---|---|---|
| 4 | Which of the following is a valid set of rules for the decision tree below?<br>  <br><br>a. IF Business Appointment = No & Temp above 70 = No **THEN** Decision = wear slacks<br>b. IF Business Appointment = Yes & Temp above 70 = Yes **THEN** Decision = wear shorts<br>c. IF Temp above 70 = No **THEN** Decision = wear shorts<br>d. IF Business Appointment= No & Temp above 70 = No **THEN** Decision = wear jeans | ( ) |
| 5 | Which of the following describes data used to build a data mining model?<br>a. Validation data<br>b. Training data<br>c. Test data<br>d. Hidden data | ( ) |

**Question 1 (25 marks)**

1 A data science company wants to hire data scientists among working adults who successfully pass the courses conducted by the company. The company wants to identify the group of candidates whom will stay on with the current company and not look for a job change. This retention helps to reduce the cost and time for training. Information related to demographics, education, and experience are gathered in a CSV file, **EST_Data_Q1.csv**.

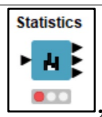
| Attribute | Description |
|---|---|
| enrolment_id | Unique ID for candidate |
| gender | Gender of candidate |
| relevent_experience | Relevant experience of candidate |
| enrolled_university | Type of University course enrolled if any |
| education_level | Education level of candidate |
| major_discipline | Education major discipline of candidate |
| experience | Candidate total working experience in years |
| last_new_job | Difference in years between previous job and current job |
| training_hours | Training hours completed |
| target | 0 – Not looking for a job change,<br>1 – Looking for a job change |

**Data file used: EST_Data_Q1.csv**

a (i) What is the size of the data? — 2 marks

| | |
|---|---|
| **Number of rows** | |
| **Number of columns** | |

a (ii) How many classes are there in the target variable? — 1 mark

b (i) Using 'Statistics' node 

, how many attributes contain missing values? — 2 marks

b (ii) Attributes will be dropped and will not be used for the model, if more than 19% (>19%) of the records are with missing values. Which attribute(s) will be dropped? — 4 marks

<table>
<tr><td>c (i)</td><td>The company wants to find out some numerical summaries of the training hours completed by all the participants. Fill in the blanks.<br><br>
<table>
<tr><th>Attribute</th><th>Median</th><th>Standard Deviation</th></tr>
<tr><td>training_hours</td><td>(A)</td><td>(B)</td></tr>
</table>
<br>
<table>
<tr><td>(A)</td><td></td></tr>
<tr><td>(B)</td><td></td></tr>
</table></td><td>4 marks</td></tr>
<tr><td>c (ii)</td><td>The company wants to compare the median and standard deviation of the training hours completed by the group whose target is not looking for job change and the group whose target is looking for a job change. Fill in the blanks.<br><br>
<table>
<tr><th colspan="3">training_hours</th></tr>
<tr><th>Target</th><th>Median</th><th>Standard Deviation</th></tr>
<tr><td>Not looking for job change</td><td>(A)</td><td>58.565</td></tr>
<tr><td>Looking for a job change</td><td>50</td><td>(B)</td></tr>
</table>
<br>
<table>
<tr><td>(A)</td><td></td></tr>
<tr><td>(B)</td><td></td></tr>
</table></td><td>4 marks</td></tr>
<tr><td>c (iii)</td><td>How many different categories are there in the attribute, experience present in the original data?</td><td>2 marks</td></tr>
<tr><td>d</td><td>Step I: Using suitable nodes, <u>filter out</u> the rows from attribute ‘experience’ which have<br>
• experience “> 20”<br>
• experience “< 1”<br>
Step II: With the remaining records from Step I, categorise the ‘experience’ column as<br>
• < 15 – Experienced<br>
• >= 15 – Very Experienced<br>
Step III: Generate a pie chart and determine the number of records and % of samples in ‘Very Experienced’ category.<br><br>
<table>
<tr><td>Number of records after Step I</td><td></td></tr>
<tr><td>Number of records in ‘Very Experienced’</td><td></td></tr>
<tr><td>% of samples in ‘Very Experienced’</td><td></td></tr>
</table></td><td>6 marks</td></tr>
</table>

**Question 2 (20 marks)**

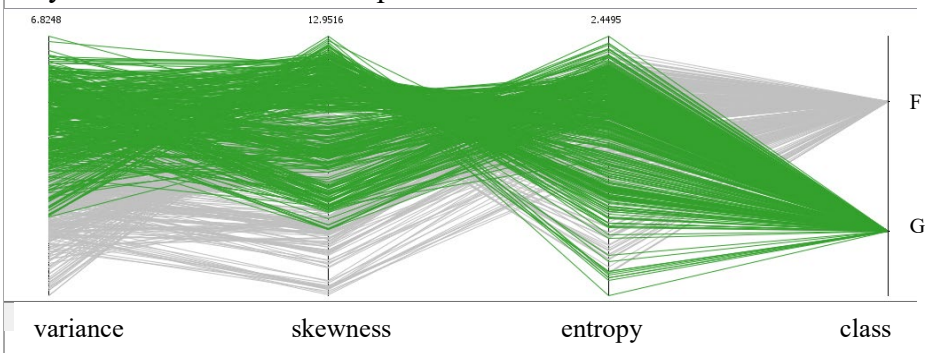
2 To develop an algorithm which can identify whether a banknote is genuine or fake, data were extracted from images that were taken from genuine (G) and fake (F) banknotes. An image processing tool was then use to extract the following variables:

| Variable | Description |
|---|---|
| **variance** | Describes how each pixel varies from the neighbouring pixels |
| **skewness** | A measure of the lack of symmetry |
| **entropy** | Amount of information which must be coded for by a compression algorithm |
| **class** | G = genuine, F = fake |

An analyst would like to use K-Means clustering to study the characteristics of the notes.

a What is the value of $k$ to be used for K-Means clustering? Briefly explain. 2 marks

b A parallel coordinates plot of the data is given below. Comment on any attributes which can help characterize the clusters. 2 marks



| | | |
|---|---|---|
| c | Explain why normalization is important in K-Means clustering. | 2 marks |
| d | A small subset of the data is given below. Use Min-max normalization to fill in the blanks (A), (B), (C) and (D) below. | 4 marks |

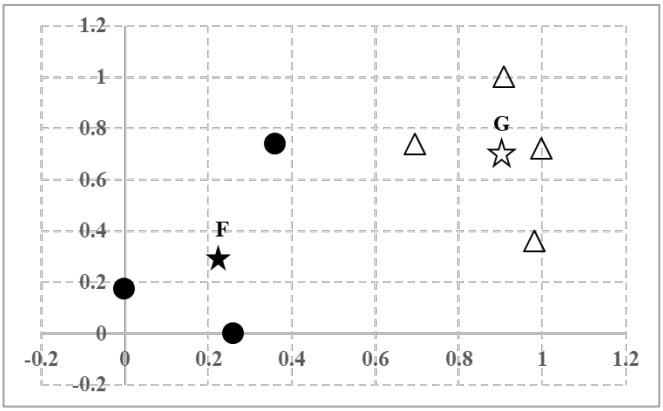
Min-Max Normalisation

| ID | Variance | Skewness | Class |
|---|---|---|---|
| **1** | 1.635 | 3.286 | G |
| **2** | 3.23 | 7.838 | G |
| **3** | 3.912 | 2.974 | G |
| **4** | 3.78 | -3.311 | G |
| **5** | -1.6 | -9.583 | F |
| **6** | -3.59 | -6.572 | F |
| **7** | -0.878 | 3.257 | F |

| ID | Variance | Skewness | Class |
|---|---|---|---|
| **1** | **(B)** | 0.739 | G |
| **2** | 0.909 | 1 | G |
| **3** | 1 | 0.721 | G |
| **4** | 0.982 | 0.360 | G |
| **5** | 0.262 | **(D)** | F |
| **6** | 0 | 0.173 | F |
| **7** | 0.361 | 0.737 | F |

| | | |
|---|---|---|
| (i) | $\frac{1.635-(-3.59)}{(\mathbf{A})}=$ | **(B)** |
| (ii) | $\frac{-9.583-(\mathbf{C})}{17.421}=$ | **(D)** |

| | |
|---|---|
| **(A)** | |
| **(B)** | |
| **(C)** | |
| **(D)** | |

| | | |
|---|---|---|
| e | A scatterplot of the data in part (d) above with two clusters is given below. The cluster centroid, F is (0.208, 0.303). Write down the cluster centroid for cluster G. Show your workings clearly. (Hint: Refer to table in part (d) above)<br>  <br> | 4 marks |
| f | Suppose a new note has the measurements variance = 2.20 and skewness = 6.00 (before standardization). Compute the Euclidean distance of the new note from each of the centroids, F and G. Which cluster is the new note likely to belong to? Explain and show your workings clearly. | 6 marks |

**Question 3 (25 marks)**

3 United Finance is a company that directly connects borrowers and potential lenders/investors. The dataset (**EST_Data_Qn3.csv**) contains details of 10,000 past loan records. In this question, you will build a decision tree classification model to predict whether or not a loan provided by United Finance is likely to be a bad loan. In other words, you will use data from the United Finance to predict whether a loan will be paid off in full or the loan will possibly go into default. The descriptions of the attributes in the dataset are given as below:

| Attribute | Description |
|---|---|
| **emp_length_num** | number of years of employment |
| **home_ownership** | home_ownership status: own, mortgage, rent or other |
| **dti** | debt to income ratio |
| **purpose** | the purpose of the loan |
| **term** | the term of the loan (in months) |
| **bad_loans** | "0" – No default<br>"1" – Default (bad debt) |

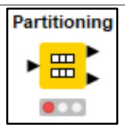
Build a decision tree classification model to predict whether a loan will go into default. Use the following setting:

i. *Size of first partition as relative 80% with stratified sampling with the 'home_ownership" attribute, using the random seed of '12345'.*
ii. *'Gini index' with no pruning.*
iii. *Reduced error pruning is checked.*
iv. *Restrict the minimum number of records per node to be 10.*

**NOTE**: Remember to change the data type of the 'bad loans' attribute from 'Integer' to 'String'

**Data file used: EST_Data_Qn3.csv**

| | | |
|---|---|---|
| a | State the target variable in this scenario. | 1 mark |
| b | Explain why a regression model is not used here. | 2 marks |
| c | Why do we need to change the data type of 'bad loans' to String? | 2 marks |

| | | |
|---|---|---|
| d | Based on the  node, what is the meaning of "first partition as relative 80%"? | 2 marks |
| e | State the quality measure that is used to build the decision tree. | 2 marks |
| f | What is the attribute for the first split of the decision tree? Briefly explain how this first split is derived. | 4 marks |
| g | State the overall accuracy of the decision tree. Show how you can use the confusion matrix generated to derive this accuracy. | 3 marks |
| h<br>(i) | From the confusion matrix generated,<br>how many false positive(s) for outcome "1" is/are produced?<br>Interpret your answer in this context. | 3 marks |
| (ii) | how many false negative(s) for outcome "1" is/are produced?<br>Interpret your answer in this context. | 3 marks |
| (iii) | If the model was deployed to predict whether to lend money to borrowers based on whether they will default, what are the implications of committing false negative errors? | 3 marks |

**Question 4 (20 marks)**

4 Prediction of property prices is becoming increasingly important and beneficial. A data analyst was tasked to analyse and predict where property prices are moving towards. After performing a regression modelling to predict the house price, the results obtained are shown below.

**Coefficients of Model 1:**

| Variable | Coeff. | P-Value |
|---|---|---|
| **Constant** | -120018 | 0.000 |
| **floors** | 45392 | 0.000 |
| **sqft_living** | 319.94 | 0.000 |
| **sqft_above** | -35.1 | 0.002 |
| **bedrooms** | -62829 | 0.000 |
| **condition** | 57093 | 0.000 |
| **sqft_lot** | -0.672 | 0.000 |

**Model Summary of Model 1:**

| R-sq | R-sq(adj) |
|---|---|
| **48.45%** | 48.38% |

| | | |
|---|---|---|
| a | State the response variable in this scenario. | 1 mark |
| b (i) | Interpret the regression coefficient of the predictor variable, **bedrooms**, of the above **Model 1**. | 2 marks |
| b (ii) | Construct the least squared regression equation for the above **Model 1**. | 4 marks |
| b (iii) | Assess whether the above **Model 1** suffers from overfitting. Why or why not? | 2 marks |

<table>
<tr><td>c</td><td>If the analyst wants to drop any variable due to insignificance, which independent variable will be eliminated for model re-building?</td><td>2 marks</td></tr>
<tr><td>d (i)</td><td>What is the coefficient of determination?</td><td>1 mark</td></tr>
<tr><td>d (ii)</td><td>Interpret the coefficient of determination.</td><td>2 marks</td></tr>
<tr><td></td><td>After a few attempts of creating the model, the data analyst added interaction terms to a regression model so as to better understand relationships among the variables in the model and allow more hypotheses to be tested. He also eliminated some features via backward elimination process. The results after adding the interaction terms are below:<br>Coefficients of Model 2:
<table>
<tr><th>Variable</th><th>Coeff.</th><th>P-Value</th></tr>
<tr><td>Constant</td><td>-122285</td><td>0.000</td></tr>
<tr><td>bedrooms</td><td>-60056</td><td>0.000</td></tr>
<tr><td>sqft_lot</td><td>-0.538</td><td>0.281</td></tr>
<tr><td>condition</td><td>53983</td><td>0.000</td></tr>
<tr><td>sqft_lot*condition</td><td>0.222</td><td>0.106</td></tr>
<tr><td>bedrooms*sqft_lot</td><td>-0.259</td><td>0.012</td></tr>
</table>
<br>Model Summary of Model 2:
<table>
<tr><th>R-sq</th><th>R-sq(adj)</th></tr>
<tr><td>48.53%</td><td>48.38%</td></tr>
</table></td><td></td></tr>
<tr><td>e</td><td>Which one of the interaction terms is insignificant and why?</td><td>2 marks</td></tr>
<tr><td>f</td><td>Which nodes would you use if you want to develop Model 2 using KNIME for the following steps?
<table>
<tr><th>Steps</th><th>Name of the KNIME Node</th></tr>
<tr><td>Create the interaction term</td><td></td></tr>
<tr><td>To evaluate and score the model</td><td></td></tr>
</table></td><td>4 marks</td></tr>
</table>