PAPUA NEW GUINEA UNIVERSITY OF TECHNOLOGY

DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE

**SECOND SEMESTER EXAMINATIONS – 2021**

**THIRD YEAR BACHELOR OF SCIENCE IN COMPUTER SCIENCE**

**CS325 – INTRODUCTION TO DATA SCIENCE**

**TIME ALLOWED: 3 HOURS**

**INFORMATION FOR CANDIDATES**

1. Write your name and student ID number clearly on the front of the examination answer booklet.

2. You have 10 minutes to read this paper. You must not begin writing during this time.

3. There are two sections in this paper. Section A contains twenty (20) multiple choice questions. Section B contains five (5) short answer questions worth eighty (80) marks. You should attempt ALL questions.

4. All answers must be written in the examination answer booklets provided. No other written material will be accepted.

5. Write out the question number clearly on the answer sheet beside each answers. Do not use red ink or pencil.

6. Notes and textbooks are not allowed in the examination room.

7. Mobile phones and other recording devices are not allowed in the examination room.

8. Scientific and Business Calculators are allowed in the exam room.

**MARKING SCHEME**

**Marks are indicated at the beginning of each question. The total is 100 Marks.**

**SECTION A – Questions 1 to 20 are multiple choice questions. Choose only one answer. You must answer all of them.**

**[1 MARK EACH = 20 MARKS]**

1. Which statement is not true about Python? Python is
   (a) an object-oriented programming language.
   (b) a procedural-oriented programming language.
   (c) emphasizes use of data structures.
   (d) a cross platform language.

2. To create a *Function* in Python, you will require two keywords. They are
   (a) Def and Return.
   (b) Declare and Return.
   (c) Def and Parm.
   (d) Def and Declare.

3. What background skills or knowledge do you need to *extract insight from data*?
   (a) Programming skills.
   (b) Maths & Statistics knowledge.
   (c) Domain knowledge.
   (d) Hacking skills.

4. Analytical 'story-telling' is an art Professionals need to develop in their professional careers. Data Analysts demonstrate their 'story-telling' art during
   (a) Data visualization and communication.
   (b) Data exploration and communication.
   (c) Data cleaning and preparation.
   (d) Data reporting & documentation.

5. Which activity is not an example of application of Big Data analytics?
   (a) Crime detection and prevention.
   (b) Optimize election campaigns.
   (c) Reporting from SQL database.
   (d) Understanding customer buying habits.

6. Which statement is not true about Logistic Regression?
   (a) Line of best fit is S-shaped.
   (b) Output values of a target variable is ordinal.
   (c) Best line is called the 'Decision Boundary'.
   (d) Its equation uses a logarithm function.

7. 'Euclidean distance' used by k-NN algorithm is the distance between data points. The method measures the
   (a) Minimum distance.
   (b) Maximum distance.
   (c) Average distance.
   (d) Median distance.

8. Given two predictor variables 'Age' and 'Education level' are used to predict 'Income level' of a person, which is *not* an example of *feature extraction* on 'Age' variable?
   (a) Calculate its logarithm.
   (b) Derive its square root.
   (c) Exclude it from modeling
   (d) Concatenate the two predictor variables.

9. Which data structure is commonly used by most programs in implementing their algorithms?
   (a) Tuples
   (b) Lists
   (c) Dictionaries
   (d) Arrays

10. Given 3-D array ( [ [[1, 2, 3], [4, 5, 6]], [[7, 8, 9] ,[10, 11, 12]] ]). Which slicing method will access element 6?
   (a) [0, 1, 1]
   (b) [0, 1, 2]
   (c) [1, 1, 1]
   (d) [1, 1, 2]

11. Which statement on Histogram is not correct? Histogram
   (a) Is a graph.
   (b) Displays number of occurrences.
   (c) Uses bin intervals
   (d) Is a statistical software application.

12. Which feature of a student is not an example of a continuous random variable?
   (a) Height
   (b) Weight
   (c) Phone numbers
   (d) Internal marks

13. Calculation of conditional probability uses mathematical expression in the form $P(Y|Z)$. How is it interpreted?
    (a) Probability of Z, given Y.
    (b) Probability of Y and Z.
    (c) Probability of Y, given Z.
    (d) Probability of Y union Z.

14. Hypothesis testing in Statistics is a type of
    (a) Alternate hypothesis
    (b) Statistical Estimation
    (c) Statistical Inferences
    (d) Null hypothesis

15. Pandas in python reads raw datasets and returns the two codes 'NaN' and 'NaT'. What do they represent?
    (a) Date stamp
    (b) Empty cell
    (c) Bad data
    (d) Time stamp

16. There are pre-requisites that are necessary before any serious deep analysis & exploration is performed on raw datasets. Which activity is <u>not</u> recommended as one of the pre-requisites?
    (a) Understand the data source
    (b) Obtain summary statistics
    (c) Perform data visualization
    (d) Clean the dataset

17. Decision tree modeling uses Information Gain (IG) to make decisions. A good model will have high IG and low Entropy in the dataset. What does Entropy measure in the dataset?
    (a) Predictions
    (b) Certainties
    (c) Uncertainties
    (d) Estimations

18. Misuse or tampering with data analysis to produce statistically significant results, when in fact there is no real underlying effect is referred to as
    (a) H-hacking
    (b) N-hacking
    (c) P-hacking
    (d) S-hacking

19. In Linear algebra, a vector
    (a) is a single element
    (b) is either a row or a column
    (c) consists of both rows and columns
    (d) consists of multiple elements

20. Which statement about an outlier or anomaly is not correct?
    (a) A potential new discovery.
    (b) Shows presence of underlying discrepancies.
    (c) Fails to reject null hypothesis.
    (d) Shows faults in measurements.

**SECTION B – Short Answer Questions**

**QUESTION 21.** **[ 3 + 5 + 2 + 3 + 2 = 15 marks ]**

a) What is *Data Science*? Provide brief explanation of its overview and what the application of Data Science entails.

b) What is *Big Data*? Provide brief explanation in terms of its five (5) characteristics, known as the 5 V's.

c) The 6$^{th}$ 'V' of Big Data is "*Value*". Briefly explain its characteristics.

d) Provide two real-world examples or applications of Big Data. Explain the examples in terms of the characteristic(s) of Big Data.

e) The phrase "*Big Data is the OIL of the new century*" is common in the industry. Briefly elaborate on it in your own words.

**QUESTION 22.** **[ (2 + 3 + 3 ) + (1 + 1 + 4 + 1) = 15 Marks ]**

a) Below is the contingency table from a dataset containing tally of students (UPNG & UoT) enrolled in respective faculties of study in 2021.

Study the table and answer the questions.

| | Faculty enrolled | | | |
|---|---|---|---|---|
| students | STEM | Business | Others | total |
| UoT | 60 | 110 | 90 | 260 |
| UPNG | 50 | 90 | 80 | 220 |
| | 110 | 200 | 170 | 480 |

(Hint: Conditional probability, P(B|A) = P(A *and* B) / P(A) where A and B are events or outcomes)

If we are to choose at random from the dataset:

i. What is the probability of selecting a student from the population who is enrolled in the *STEM* program? Use appropriate notations in your calculations.

ii. What is the probability of selecting a student at *UoT*, given program preference is *STEM*? Use appropriate notations in your calculations.

iii. What is the probability of selecting a student enrolled in the Business faculty, who is studying at UPNG? Use appropriate notations in your calculations.

b) Given the sample data values (1, 2, 5, 7, 8, 10), answer the following questions.

i. What is the total number of observations
ii. Calculate the sample mean
iii. Calculate the standard deviation (Use the formula provided)
iv. Show your calculations and workouts.

**Sample**

$$s = \sqrt{\frac{\Sigma(X - \bar{x})^2}{n - 1}}$$

*$X$ – The Value in the data distribution*
*$\bar{x}$ – The Sample Mean*
*$n$ - Total Number of Observations*

**QUESTION 23.** **[ (6 + 2 + 2) + 5 + 5 = 20 Marks ]**

a) Table below shows 4 short text messages used by a Naïve Bayes classifier to label whether a text message is 'Junk' or 'Genuine'. A new text "Please, call me" is to be labeled accordingly.

| **SMS text message** | **Tag** |
|---|---|
| *Call me now* | *Genuine* |
| *Call now asap* | *Junk* |
| *Call me asap* | *Genuine* |
| *Call asap and win* | *Junk* |
| **New text** | |
| *Please call me* | *???* |

Calculate its probability score and label whether it is 'Junk' or 'Genuine'.

i. Prepare word frequency or word-count table from data
ii. Calculate probability score for the new text message
iii. Label the new text message

b) Select the correct modeling type (algorithm) from the list and write on your answer sheet.

- SVM; k-NN; Logistic regression; Clustering; Decision Tree; Linear Regression; Naïve Bayes

| **Type of model** | **Brief meaning/description** |
|---|---|
| i. | Assumes that data points that are far away from each other are dissimilar. |
| Ii. | 1. Consumes more time in training/testing, not suitable for large datasets.<br>2. Does not work well with overlapping classes |
| iii. | Right questions need to be asked for optimal performance |
| iv. | Line of best fit is not linear but curve-shaped. |
| v. | Application of this model -> Air Niugini offers discounts to tertiary students travelling after the exams. |

c) Select the correct term from the list and write on your answer sheet.

- Matplotlib; Correlation; Array; Histogram; Scatterplot; Numpy; Pandas

| **Concept/term** | **Brief meaning/description** |
|---|---|
| i. | Data structure for numerical values |
| ii. | Uses DataFrame |
| iii. | Shows a tally on Frequency Counts |
| iv. | Numerical comparison between predicting pairs |
| v | Data structure used by most programming algorithms |

**QUESTION 24.** **[ 3 + 4 + 4 + 1 + 1 + 2 = 15 Marks ]**

A 'Multi-linear Regression' modeling was performed on a dataset from an Auto-mobile dealer in Lae who operates a fleet of new and old vehicles. The dataset has the flowing attributes:

- Repair-cost: Maintenance/repair cost of the vehicle in kina (1 unit = K100)
- Age: Age of vehicle in years (from year of manufacture)
- Space: Space capacity measured in square meters (sqm)
- Mileage: Accumulated distance travelled by the vehicle, in kilometres (km)
- Vehicle-value: Economical value of the vehicle in kina

Study the summary output below and answer the questions.

Summary output - Regression Statistics

| | | |
|---|---|---|
| *R-square* | *0.78* | |
| *Observations* | *20* | |
| | | |
| ***Predictor variables*** | ***Coefficients*** | ***P-value*** |
| *Intercept* | *10,000* | *0.0023* |
| *RepairCost* | *150* | *<0.0001* |
| *Mileage* | *-50* | *<0.0001* |
| *Age* | *-85* | *<0.0001* |
| *Space* | *30* | *0.900* |
| | | |
| *Responsive variable:* | *Vehicle-value* | |

a) Construct a Multiple-linear regression equation

b) Interpret the Multiple linear regression equation

c) Given cut-off mark for p-value is **0.05,** interpret *p-values* for the predictor variables.

d) Interpret the *R-square* of the model

e) Interpret *Observations* on the table

f) List down one factor (predictor variable) you think can be added to the model to improve prediction, and why do you think so?

**QUESTION 25** **[ (1 + 1 + 1 + 4) + (1 + 1) + 2 + 2 + 2 = 15 marks ]**

The BSP bank in POM implemented a modeling algorithm to predict likelihood of their customers leaving BSP and switching over to the new Kina Bank. The model was trained on 1,000 customer cases to predict whether the customer is 'leaving' or 'not leaving'.
BSP is interested to know in advance which customers are leaving.

Outcomes from modeling 1,000 cases are as follows;

- 40 cases, prediction is TRUE, actually FALSE
- 50 cases, prediction is FALSE, actually TRUE
- 110 cases, prediction is TRUE, actually TRUE
- 800 cases, prediction is FALSE, actually FALSE

*( Hint: FP – false positive; FN – false negative; TN – true negative; TP – true positive )*

Study the output from the model below and answer the questions.

a) Construct a confusion matrix table with the above information
   i. Label with 'Actual values' and 'Predicted values'
   ii. Label with 'Positive' and 'Negative'
   iii. Label with 'Leaving' and 'Not leaving'
   iv. Allocate cases into correct buckets of FP, FN, TN, TP

b) Calculate the Accuracy and F1 score of the model (in %), where;

   i. Accuracy = (TP + TN) / (TP + FP + FN + TN)

   ii. F1 score =( 2TP) / (2TP + FP + FN)

c) Identify *type1* error from the table and explain in relation to the outcome

d) Identify *type2* error from the table and explain in relation to the outcome

e) Once the bank has correctly identified their customers who are most likely to 'leave'. What do you think the bank should do to these customers? Explain in terms of the application of Big Data analytics.

**END OF EXAM**