THE PAPUA NEW GUINEA UNIVERSITY OF TECHNOLOGY

DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE

SECOND SEMESTER EXAMINATIONS – 2022

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 number clearly on the front of the answer booklet.
2. You have 10 minutes to read this paper. You must not begin writing during this time.
3. Section A has **20** multiple choice questions and Section B has **five** questions. You should answer all the questions.
4. All answers must be written in examination answer booklets only. No other written material will be accepted.
5. Write all the answers for Section A on one page. For Section B, start the answer for each question on a **new** page. Do **not** use red ink.
6. Notes and textbooks are not allowed in the examination room. All mobile phones and electronic/recording devices must be switched off during the examination.
7. Scientific and business calculators are allowed in the examination room.

MARKING SCHEME

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

SECTION A **[1 mark each = 20 marks]**

Choose the correct answer and write A, B, C or D from the alternatives given.

1. Which one of these statements is not true about programming syntax for Python?
   A. A variable is declared first before it is used.
   B. Uses hash (#) symbol as a comment line.
   C. Reserve keywords are case-sensitive.
   D. Values of the variables are defined on the right side of the '=' sign.

2. Certain keywords are used to define specific objects in Python. Which one of these objects is defined by *def* and *return* keywords?
   A. Function.
   B. Procedure.
   C. Sub-routine.
   D. Parameter.

3. Data scientists and statisticians are professionals in data analytics. Which one of these skill-sets and background differentiates a *data scientist* from a *statistician*?
   A. Probability.
   B. Predictive Analytics.
   C. Programming.
   D. Statistical Inferences.

4. Python uses a specific library that contains functionality to explore and communicate data using data-driven visuals. Which one of these libraries enables *data visualization* in Python?
   A. Matplotlib.
   B. Pyplotlib.
   C. Scatterplotlib.
   D. Ggplotlib.

5. Real-world dataset contains *bad data* which are typically cleansed during the pre-processing phase. Which one of these examples is considered as *bad data*?
   A. Invalid information.
   B. Data entry error.
   C. Result from incorrect formula.
   D. Wrong format.

6. A *decision-tree* algorithm has two main entities called *decision-node* and *leaf*. Which one of these features of a dataset corresponds to *decision-node* and *leaf* respectively?
   A. Predictor variable, input variable.
   B. Predictor variable, outcome variable.
   C. Outcome variable, predictor variable.
   D. Outcome variable, input variable.

7. Performance of a *decision-tree* algorithm is at its optimum level if there is high *Information Gain (IG)* and low *Entropy* in the dataset. Which of these statements about *IG* and *Entropy* is not relevant?
   A. IG reduces entropy.
   B. Entropy is the measure of uncertainty in the data.
   C. High entropy means low certainty.
   D. None of the above.

8. *Clustering* and *k-NN* algorithms classify data points as *outliers* basing on certain conditions. Which one of these basis is a condition for an *outlier*?
   A. Minimum distance from its nearest neighbor.
   B. Maximum distance from its nearest neighbor.
   C. Average distance from its nearest neighbor.
   D. Median distance from its nearest neighbor.

9. Modelers apply *feature selection* and/or *feature extraction* to choose optimum features for modeling. Which one of these statements is not correct?
   A. Feature extraction is based on original attributes.
   B. Feature extraction uses a subset of the original attributes.
   C. Feature extraction derives a new attribute from the original attribute.
   D. Feature extraction reduces model complexity.

10. An array is a data structure implemented by most programs in their algorithms. Matrix is one special type of array that exists as a
   A. 0-D array.
   B. 2-D array.
   C. 3-D array.
   D. n-D array.

11. If a 3-D array is [[[1, 2, 3], [4, 5, 6]], [[7, 8, 9] ,[10, 11, 12]]] , which slicing method will access element **10**?
   A. [0, 1, 1]
   B. [0, 1, 2]
   C. [1, 1, 0]
   D. [1, 1, 1]

12. A *histogram* shows data distribution of an attribute of interest in a dataset. To manually create a simple histogram, what is the next step taken after the raw data is analyzed?
   A. Sort data in descending order.
   B. Count and tally the number of occurrences.
   C. Create number of bin intervals.
   D. Sort data in ascending order.

13. Which one of these attributes of a PMV bus is an example of a *discrete* variable?
    A. Mileage.
    B. Tonnage.
    C. Seating capacity.
    D. Spacing capacity.

14. The shape of a *normal distribution* curve is directly influenced by two statistical properties of the dataset. Which one of these two properties can easily influence the distribution?
    A. Mean and quartile.
    B. Mean and median.
    C. Mean and standard deviation.
    D. Mean and frequency.

15. One of the statistical outputs of modeling is the *p-value. P-value* is a measure of the probability that the changes in the observed outcome variable
    A. Occurs by a random chance.
    B. Is statistically significant.
    C. Supports the claim by alternate hypothesis.
    D. Is not statistically significant.

16. *Pandas* in Python inserts a certain character string to represent a blank value for a *date* attribute. Which one of these strings represents a *missing date* in a dataset?
    A. NaN
    B. NaT
    C. Null
    D. NaD

17. The *knowledge discovery* process converts raw data into useful information and insights. Which one of these activities is the initial step in *knowledge discovery*?
    A. Data visualization.
    B. Data communication.
    C. Data modeling.
    D. Data exploration.

18. Tampering with or misuse of data analysis to produce statistically significant results, when in fact there is no real underlying effect is referred to as:
    A. Mis-classification.
    B. Plagiarism.
    C. P-hacking.
    D. Cherry-picking.

19. In *linear algebra*, individual elements are stored in data structures such as lists, tuples, arrays and vectors. A *vector* is different from the other structures because it
    A. Performs mathematical operations
    B. Handles numerical data types
    C. Occupies a finite dimensional space
    D. Is a special type of data structure

20. In *machine learning*, a new data object is considered a new discovery potential if its characteristics
    A. Do not conform to normal expected pattern.
    B. Show improvements in the prediction rate.
    C. Conform to normal expected pattern.
    D. Are similar to existing members.

SECTION B – Short Answer Questions

QUESTION 21. **[(2 + 2 + 2 + 2) + ((3 + 3 + 2 + 2) or (4 + 6)) = 18 marks]**

(a) Data science & Big data.

(i.) Briefly explain *Data Science* in terms of its overview and significance in the industries.

(ii.) Briefly explain *Big Data* in terms of its characteristics and the big Vs.

(iii.) Choose two (2) of the big Vs, and provide a real-world example or data source for each of them.

(iv.) One of the big Vs of big data is *Value*. Briefly explain its characteristics and its importance in the industries.

Choose either question (b) or (c) from below and answer the questions

(b) Data mining and Machine learning.

(i.) Briefly explain *data mining* in relation to *big data.*

(ii.) Briefly explain *machine learning* in relation to *big data.*

(iii.) List <u>two</u> similarities between the two concepts.

(iv.) List <u>two</u> differences between the two concepts.

(c) According to the CRISP-DM cycle, *data mining* process consists of six (6) phases: *Business understanding; Evaluation; Data understanding; Deployment; Data preparation; Modeling*

(i.) Draw the CRISP-DM cycle, and label the six (6) phases on the diagram. *(CRISP-DM is an acronym for Cross-Industry Standard Process for Data Mining)*

(ii.) Choose any three (3) phases and provide brief explanations on each of them.

QUESTION 22. **[(2 + 3 + 3) + (2 + 4) = 14 Marks]**

(a) Below is the contingency table from a dataset containing PNGUoT staff from different cultural backgrounds, who are keen supporters of the NRL's State of Origin games in Australia.

Study the table below and answer the questions, using the appropriate notations.

	Citizen			
Supporter	*PNG*	*Fiji*	*NZ*	*Total*
Blues	25	10	3	38
Maroons	20	15	7	42
	45	25	10	80

(Conditional probability, $P(B|A) = \frac{P(A \text{ union } B)}{P(A)}$*, where A and B are separate events or outcomes)*

If a selection is done at random from the dataset,

(i.) What is the probability of selecting a PNG citizen?

(ii.) What is the probability of selecting a Fijian, given the Fijian is a Maroons fan?

(iii.) What is the probability of selecting a Maroons supporter, given the fan is from NZ?

(b) Given sample data values [1, 3, 7, 9 10], use the hints provided and manually calculate the *standard deviation. (Show your calculations and workouts).*

Hints: Steps to calculate standard deviation for a sample population

Step1: Calculate the MEAN
Step2: Calculate the DEVIATION from the MEAN
Step3: Square the DEVIATION
Step4: SUM the square deviations
Step5: Divide the SUM in step4 by (n-1)
Step6: Take the SQRT of the quotient in step5

Sample

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

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

QUESTION 23. **[(4 + 2 + 2) + 5 + 5 = 18 Marks]**

(a) *Naïve Bayes* classifier tagged the four (4) short text messages below as either a 'Sick' or 'Healthy' related message. A new text "*Well get rest*" is to be labeled accordingly.

SMS text message	**Class**
Feel sick	*Sick*
Need rest	*Sick*
Get some rest	*Healthy*
Get well	*Healthy*
New incoming text	
Well get rest	*?*

(i.) Prepare and show word-frequency (or word-count) table from data.
(ii.) Calculate probability score for the new text message.
(iii.) What is the class of the new text message?

(b) Select the correct *algorithm* from the list and write on your answer booklet.

SVM; k-NN; Decision-tree; Logistic regression; Clustering; Naïve Bayes

Machine Learning algorithm	Brief meaning/description
(i.)	Uses K-means algorithm based on minimum distance.
(ii.)	Classifier does not work well with over-lapping classes.
(iii.)	Splits the dataset in different ways based on different conditions.
(iv.)	Uses regression equation but outcome is dichotomous and/or discrete.
(v.)	As dimension increases, closest distance between neighboring data points approaches average, hence loses predictive power.

(c) Select the correct *Python* object from the list and write on your answer booklet.

Math; Pyplot; PyStats; CSV; Numpy; Pandas

Python library/function	Functionality
(i.)	Manipulates arrays of data.
(ii.)	Converts dictionaries into tables.
(iii.)	Creates data visuals.
(iv.)	Loads external data.
(v.)	Enables auto calculations.

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

A *multiple linear regression* modeling was performed on a weekly sales volume of data from SVS Supermarket in the 2021 financial year. File definition for the *sales* table is:

Attribute	**Description**	**Unit of measure**
Sales	Revenue from sales.	Kina
Ad-cost	Cost of advertising.	Kina
Opex	Cost of general operating expenses.	Kina
Security	Number of security guards employed.	Integer
Profit	Net profit from sales.	Kina

Study this summary output (regression statistics) and answer the questions.

R-square	0.75	
Observations	52	
Predictor variables	**Coefficients**	**P-value**
Intercept	10,000	0.0020
Sales	50	0.0005
Ad-cost	30	0.0003
Opex	40	0.0001
Security	-200	1.920
Responsive variable:	Profit	

(a) Construct a *multiple-linear regression* equation.

(b) Interpret the *coefficients* and *p-values* of the model.

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

(d) Which variable(s) should be dropped from modeling, and why? How would you interpret this technical scenario to SVS management?

(e) Interpret the *R-square* of the model?

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

Lae-town Police implemented one of the *machine learning* algorithms to predict likelihood of a *civil unrest* in order for them to take preventative measures. The model was trained on 100 cases to predict whether there will be 'Unrest' or 'No-*unrest*' on a particular day.
The aim of this implementation is to know in advance the likelihood of a civil '*unrest*'.

Outcomes from modeling 100 cases are as follows;

- 4 cases, prediction is TRUE, actually FALSE.
- 5 cases, prediction is FALSE, actually TRUE.
- 11 cases, prediction is TRUE, actually TRUE.
- 80 cases, prediction is FALSE, actually FALSE.

(Hint: FP – False Positive; FN – False Negative; TN – True Negative; TP – True Positive)

Study the outputs from the model and answer the questions.

(a) Construct a confusion matrix table with the above information.
  (i.) Label the matrix with '*Actual*' and '*Predicted*'.
  (ii.) Label the matrix with '*Positive*' and '*Negative*' or '*+ve*' and '*-ve*'.
  (iii.) Label the matrix with '*Unrest*' and '*No-unrest*'.
  (iv.) Allocate the 100 cases onto the 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) How many cases are predicted as *Type1 error*? How do you explain it?

(d) How many cases are predicted as *Type2 error*? How do you explain it?

(e) Which type of error is regarded as the 'worst error' in predictions as shown in this example? Is it type1 or type2, and why do you think so?

END OF EXAMINATION