NAMIBIA UNIVERSITY
OF SCIENCE AND TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATICS

DEPARTMENT OF INFORMATICS

QUALIFICATIONS: Bachelor of Informatics	
QUALIFICATION CODE: 06BENT/06BAIF	**LEVEL: 6**
COURSE CODE: DTA621S	**COURSE**: Data Analytics
DATE: October 2025	**SESSION:** 1
DURATION: 3 Hours	**MARKS: 100**

FIRST OPPORTUNITY EXAMINATION QUESTION PAPER	
EXAMINERS:	Dr Clopas Kwenda
MODERATOR(S):	Professor Stephen Fashoto

THIS EXAMINATION PAPER CONSISTS OF 7 PAGES

(INCLUDING THIS FRONT PAGE)

INSTRUCTIONS FOR THE CANDIDATE

1. Answer ALL QUESTIONS.
2. When writing, take into account: The style should inform than impress, it should be formal, in third person, paragraphs set out according to ideas or issues, and the paragraphs flowing in a logical order.
3. Information should be brief and accurate.

Please ensure that your writing is legible, neat and presentable

Question One

1. What is the primary purpose of the .plot.scatter() **function in Pandas?**

a) To show the distribution of a single numerical variable.

b) To visualize the relationship between two numerical variables.

c) To compare the proportion of different categories in a dataset.

d) To identify missing values in a DataFrame

2. The code data2.isnull().sum() **is used to:**

a) Remove all missing values from the DataFrame data2.

b) Replace all missing values in data2 with the value 0.

c) Count and display the number of missing values in each column of data2.

d) Create a histogram showing the frequency of missing values.

3. According to the notes, why is it generally "not advisable" to use dropna() **to remove all observations with missing values?**

a) Because the dropna() function has syntax errors and doesn't work.

b) Because it can only be used on categorical variables, not numerical ones.

c) Because it can significantly reduce the dataset size and lead to less effective or biased analysis.

d) Because it is slower than imputing the values with the mean.

4. The process of filling in missing values with estimated ones (like the mean, median, or a specified value) is known as:

a) Deletion

b) Validation

c) Imputation

d) Visualization

5. Look at the following line of code:

loan['LoanAmount'].fillna(loan['LoanAmount'].median(), inplace=True)

What is the result of executing this code?

a) It creates a median value for the 'LoanAmount' column.

b) It deletes all rows where the 'LoanAmount' value is missing.

c) It replaces all missing values in the 'LoanAmount' column with the median value of that column.

d) It counts how many missing values are in the 'LoanAmount' column.

6. The equation $Y = b_0 + b_1x + b_2x^2$ is an example of which type of regression?

a) Linear Regression

b) Polynomial Regression

c) Logistic Regression
d) Ridge Regression

7. Predicting the price of a house based on its features like size and location is a classic example of a:

a) Regression problem
b) Classification problem
c) Clustering problem
d) Dimensionality reduction problem

8. A dataset where each row represents a person and columns include 'Age' (numerical), 'Purchased' (Yes/No), and 'Country' (text) is an example of:

a) A tree-like dataset
b) A tabular dataset
c) A JSON dataset
d) An unstructured dataset

9. An agent that learns to perform a task by receiving rewards for good actions and penalties for bad ones is an example of:

a) Supervised Learning
b) Unsupervised Learning
c) Reinforcement Learning
d) Regression Analysis

10. In the machine learning life cycle, which step involves cleaning the data and handling missing values?

a) Gathering Data
b) Data Analysis
c) Data Wrangling
d) Deployment

11. A model that uses curves, jumps, or twists in math to fit messy, real-world data (like predicting viral social media posts) is called a:

a) Linear Model
b) Non-Linear Model
c) Descriptive Model
d) Deterministic Model

12. Which modeling principle involves updating probabilities as new evidence becomes available, much like a spam filter?

a) Occam's Razor

b) Seek Consensus
c) Use Bayesian Reasoning
d) Update Forecasts Dynamically

13. A weather forecast that says "70% chance of rain" is an example of a:

a) Deterministic Model
b) Stochastic Model
c) First-Principles Model
d) Linear Model

14. Which of Nate Silver's principles involves combining multiple models or data sources to improve reliability?

a) Think Probabilistically
b) Update Forecasts Dynamically
c) Seek Consensus
d) Use Bayesian Reasoning

15. A model that provides a single, exact prediction with no element of probability is called a:

a) Stochastic Model
b) Deterministic Model
c) Probabilistic Model
d) Data-Driven Model

16. Descriptive analytics mainly answers the question:

a) What will happen?
b) Why did it happen?
c) What should we do?
c) What happened?

17. A sales report showing monthly revenue trends is an example of:

a) Diagnostic analytics
b) Descriptive analytics
c) Predictive analytics
d) Prescriptive analytics

18. Which source is an example of *internal data*?

a) Public datasets
b) Web scraping
c) CRM software
d) Surveys

19. Which statement best describes semi-structured data?

a) Data always stored in SQL databases

b) Has some organizational tags but no strict schema

c) Has no format or organization at all

d) Randomized raw data

20.Which method is NOT a valid way to add a new key-value pair to a dictionary?

a) dict['new_key'] = 'new_value'

b) dict.update({'new_key': 'new_value'})

c) dict.add('new_key', 'new_value')

d) dict.setdefault('new_key', 'new_value')

Question two

1. **Elements of a tuple can be accessed using indexing and slicing, similar to lists.**
   True / False
2. **The** sorted() **function modifies the original tuple when it sorts it.**
   True / False
3. **You can create a tuple from a list using the** tuple() **function.**
   True / False
4. **You can use the** append() **method to add a new element to an existing tuple.**
   True / False
5. **The main goal of the bias-variance trade-off is to maximize both bias and variance for the most robust model.**
   True / False
6. **A model that performs well on training data but poorly on new, unseen data is likely overfitting.**
   True / False
7. **Relational operators always give the same type of output when applied to both 1-D and multi-dimensional arrays.**
   True / False
8. **The transpose() function changes the orientation of a multidimensional array, while the flatten() function collapses it into a 1-D array.**
   True / False
9. **The zeros() function creates an array filled with zeros, while the ones() function creates an array filled with ones.**
   True / False
10. **The arange() function is used to create arrays with regularly spaced values within a specified interval.**

True / False

Question three

a) With regard to supervised learning, explain the following terms

i. Regression task **(2 marks)**

ii. Classification task **(2 marks)**

iii. Multicollinearity **(2 marks)**

iv. Overfitting **(2 marks)**

b) Differentiate between the two types of "missing values" mentioned in the notes: `NA` and `NaN`. Provide an example of how a `NaN` value might be generated. **(5 marks)**

c) What does the `dropna(inplace=True)` method do to a DataFrame? **(2 marks)**

d) Define the term **imputation** in the context of data preprocessing. **(2 marks)**

e) Give two examples each for **(8 marks)**

i. Structured data
ii. Semi-structured data
iii. Unstructured data
iv. Ordinal data

Question four

a) The manager of "The Daily Grind" coffee shop wants to analyze the consistency of their morning sales. The sales revenue (in dollars) for the past week (Monday to Saturday) was recorded as follows:

Sales: 102, 115, 98, 107, 110, 108

Calculate the following statistics to summarize the spread of this data:

I. The **range**. **(2 marks)**

II. The **sample standard deviation**. **(5 marks)**

III. The **sample variance**. **(2 marks)**

IV. The **coefficient of variation (CV)**. **(2 marks)**

b) The students in Mr. Smith's Period 1 math class were asked how many hours they spent studying for their last test. Their responses, listed in order from least to greatest, are as follows:

5, 6, 6, 7, 8, 9, 10, 10, 11, 12, 15

1. Calculate the following statistics needed to create a box and whisker plot:
   i. Minimum **(1 mark)**
   ii. First quartile (Q1) **(1 mark)**
   iii. Median **(1 mark)**
   iv. Third quartile(Q3) **(1 mark)**
   v. Maximum **(1 mark)**
2. Draw a box and whisker plot that represents the data above **(Hint use a number line from 0 to 18) (5 marks)**
3. Use your box plot to answer the following questions:
   i. What is the **interquartile range (IQR)**? **(2 marks)**
   ii. Would you describe the distribution of study times as symmetrical, skewed left, or skewed right? Explain your reasoning based on the shape of the box plot. **(2 marks)**

Question five

a) A survey was conducted to evaluate a student's performance in **five skills**:

Skill	Communication	Teamwork	Problem-Solving	Creativity	Technical Skills
Score (out of 10)	8	7	9	6	8

Task:

I. Create a **radar chart (spider chart)** to visualize the student's performance across the five skills. **(10 marks)**

b) Data wrangling involves cleaning and transforming raw data into a usable format. List and explain four techniques commonly used in the data wrangling process **(6 marks)**

c) Supervised learning models can be categorized into regression and classification. Explain the difference between these two types of models. **(4 marks)**