NAMIBIA UNIVERSITY
OF SCIENCE AND TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATICS

DEPARTMENT OF INFORMATICS

QUALIFICATION : BACHELOR OF INFORMATICS, BACHELOR OF COMPUTER SCIENCE	
QUALIFICATION CODE: 07BAIT,07BCMS	**LEVEL:** 6
COURSE: DATA ANALYTICS	**COURSE CODE**: DTA621
DATE: NOVEMBER 2024	**SESSION:** 1
DURATION: 2 HOURS	**MARKS:** 85

FIRST OPPORTUNITY EXAMINATION QUESTION PAPER	
EXAMINER(S)	MRS RUUSA IPINGE
MODERATOR:	MR SEBASTIAN MUKUMBIRA

THIS QUESTION PAPER CONSISTS OF 10 PAGES

(Excluding this front page)

INSTRUCTIONS

- Answer ALL questions in Part 1, Part 2 and Part 3,
- NUST examinations rules apply.
- DO NOT open this examination cover until you are instructed to do so.
- DO NOT FORGET to write down your student number at the designated places in the examination page.

PART 1: MULTIPLE CHOICE QUESTIONS (25 MARKS MAXIMUM 1 MARK FOR EACH CORRECT ANSWER)

Answer all questions. Select ONLY ONE BEST ANSWER to each question.

1. **___ is a type of supervised machine learning methods that uses the sigmoid function) to map predicted values to probabilities between 0 and 1.**

   a) Classification
   b) Clustering
   c) Logistic Regression
   d) None of the mentioned above

2. **An advantage of using computer programs for qualitative data is that they ___.**

   a) Can reduce time required to analyse data.
   b) Help in storing and organizing data.
   c) Make many procedures available that are rarely done by hand due to time constraints.
   d) All the mentioned above

3. **Logistic regression is used to find _____ of event = Success and event = ____.**

   a) Binary
   b) Function
   c) Probability
   d) None of the mentioned above

4. **This refers to techniques used to reduce noise or fluctuations in data, making patterns more discernible.**

   a) Smoothing
   b) Data aggregation
   c) Discretization
   d) Normalisation

5. **This is the type of research that ~~It~~ answers key questions such as "how many, "what" and "why".**

   a) Quantitative
   b) Qualitative
   c) Nominal
   d) Category

6. **___ is a plot that is good when you have a small number of categories, typically less than five or six variables, whereby you want to highlight the percentage of each category in relation to the total.**

   a. Bar graph
   b. Scatterplot
   c. Line graph
   d. Pie chart

7. **This is the process reorganising data and cleaning data by removing redundant and unstructured data and making the data look similar across all records**

   a) Smoothing
   b) Data cleaning
   c) Discretization
   d) Normalisation

8. **This is the type of research that It answers key questions such as "how many, "how much" and "how often".**

   a) Quantitative
   b) Qualitative
   c) Nominal
   d) Category

9. **This is an example of ordinal data:**
   a) The amount of time required to complete a project.
   b) The weight of children.
   c) The square footage of a two-bedroom house.
   d) The number of books read by students in a month

10. **Which statement is true about ordinal data**
   a) You cannot do arithmetic with ordinal numbers because they only show sequence.
   b) Ordinal variables are considered as "in between" qualitative and quantitative data
   c) The ordinal data is qualitative data for which the values are ordered.
   d) All the mentioned above

11. **These refer to provision of functionalities like cross-validation, hyperparameter tuning, and performance metrics, helping in making the selection process more efficient and robust**

   a) Model Selection
   b) Training Data set
   c) Testing Data set
   d) Supervised Machine learning

12. **Amongst which of the following is / are the applications of Linear Regression,**

   a) Biological
   b) Behavioural
   c) Social sciences
   d) All the mentioned about

13. **Refers to the different types and sources of data that organizations collect and analyse**

   a) Value
   b) Variety
   c) Velocity
   d) None of the mentioned above

14. **This is the process of designing, building, and maintaining systems that collect, store, and process large volumes of data.**

a) Data mining
b) Data Engineering
c) Data warehouse
d) All of the mentioned above

15. **In Shayla's math class, she asks eight people out of the forty people in the class what grade they earned on the last exam. The data she collected is shown below. What is the sample mean for this sample?**

Test scores: 89, 100 61, 100, 95, 76, 83, 91

a) 81.5

b) 84.2

c) 78.3

d) 86.8

16. **You want to identify global weather patterns that may have been affected by climate change. To do so, you want to use machine learning algorithms to find patterns that would otherwise be imperceptible to a human meteorologist. What is the place to start?**

a) Find labelled data of sunny days so that the machine will learn to identify bad weather.
b) Use unsupervised learning have the machine look for anomalies in a massive weather database.
c) Create a training set of unusual patterns and ask the machine learning algorithms to classify them.
d) Create a training set of normal weather and have the machine look for similar patterns

17. **What is one reason not to use the same data for both your training set and your testing set?**

a) You will almost certainly underfit the model.
b) You will pick the wrong algorithm.
c) You might not have enough data for both.
d) You will almost certainly overfit the model.

18. **What is the primary objective of the GDPR?**

a) To promote online marketing
b) To protect the fundamental rights and freedoms of individuals
c) To restrict international data transfers
d) To enforce mandatory data retention policies

19. **What is the definition of personal data under the GDPR?**

a) Only sensitive information
b) Any information related to an identified or identifiable natural person
c) Business-related data
d) Publicly available information

20. **Which lawful basis for processing personal data requires explicit, informed consent?**

a) Legitimate interests
b) Contractual necessity
c) Vital interests
d) Consent

21. **What is the function used to group data by one or more columns in Pandas?**

a) df.group()
b) df.aggregate()
c) df.groupby()
d) df.partition()

22. **How can you drop a column named 'age' from a DataFrame df?**

a) df.remove('age')
b) df.drop('age', axis=1)
c) df.delete('age')
d) df.pop('age')

23. **Which method would you use to fill missing values in a DataFrame?**

a) df.fillna()

b) df.replace_na()

c) df.impute()

d) df.na.fill()

24. **How can you find the shape of a NumPy array arr?**

a) arr.size

b) arr.shape()

c) arr.shape

d) np.shape(arr)

25. **What is the primary data structure used in NumPy?**

a) List

b) Dictionary

c) Array

d) DataFrame

PART 2: STRUCTURED QUESTIONS

ANSWER ALL QUESTIONS

QUESTIONS 1

1. Explain the difference between the following term [10]
   a) Supervised and unsupervised machine learning
   b) Training and testing Database
   c) Linear and Multiple regression
   d) Underfitting and Overfitting
   e) Variance and Standard Deviation

QUESTION 2

2. A class contains 39 children. The following children were chosen at random, and their weight were recorded in cm: 25, 26, 27, 30, and 32. Calculate their age. Calculate the variance of their age. Show your work [5]

3. What is Accuracy Score, what does it measure? [2]

QUESTION 3

4. Explain the output of the following python codes [10]

a)
```
x = 6
y = 7
print(type(x))
```

b)
```
d =77
r= 88
if d > r:
  print("r is greater than a"
else
  print(" No output")
```

c)
```
names = ("Selma", "Ruusa", "Suama","Thomas"))
print(type(names))
```

d)
```
set1 = {"n", "k" , "l"}
set2 = {"n", "k" , "l"}
set3 = set1.union(set2)
print(set3)
```

e)
```
def my_function(*kids):
  print("The youngest child is " + kids[2]
```

QUESTION 4

5. Describe the output of the following code: [10]

a)
```
tt = GaussianNB()
    tt.fit(X_train, Y_train)
```

b)
```
df.hist()
    plt.show()
```

c)
```
df.drop(["class"],axis=1)
```

d)
```
df.shape()
```

e)
```
df.describe()
```

6. Using the following confusion matrix. Calculate the following and interpreter with really example of what the results mean. Show your work [9]

Predicted without the Variant	**Predicted with the Variant**
TP=200	FP=150
FN=105	TN=45

i. Recall rate.
ii. Accuracy
iii. Specificality

QUESTION 5

a) Explain the 7 fundamental rights of the General Data Protection Regulation? [14]

END OF QUESTION PAPER