National University of Computer and Emerging Sciences

FAST School of Computing **Fall-2023** **Islamabad Campus**

DS2001: Introduction to Data Science

Serial No:

2nd Sessional Exam

Total Time: 1 Hour

Total Marks: 55

Saturday, 4th November, 2023

Course Instructors

Dr. Ramoza Ahsan, Bushra Amjad, Khadija Mahmood

____

Signature of Invigilator

____________	____________	____________	____________
Student Name	**Roll No.**	**Course Section**	**Student Signature**

DO NOT OPEN THE QUESTION BOOK OR START UNTIL INSTRUCTED.

Instructions:

1. Attempt on question paper. Attempt all of them. Read the question carefully, understand the question, and then attempt it.
2. No additional sheet will be provided for rough work. Use the back of the last page for rough work.
3. If you need more space, write on the back side of the paper and clearly mark question and part number etc.
4. After asked to commence the exam, please verify that you have **Fourteen (14)** different printed pages including this title page. There are total of **4** questions.
5. Calculator sharing is strictly prohibited.
6. Use permanent ink pens only. Any part done using soft pencil will not be marked and cannot be claimed for rechecking.
7. If you have read the instructions, circle your instructor's name to get 2 bonus marks.

	Q-1	Q-2	Q-3	Q-4	Total
Marks Obtained					
Total Marks	20	15	10	10	55

Roll No.

For Question 1 (MCQs), mark all the answers on this sheet. Any MCQ mark other than this sheet will not result in any marks.

A B ● D

1. A B C D

2. A B C D

3. A B C D

4. A B C D

5. A B C D

6. A B C D

7. A B C D

8. A B C D

9. A B C D

10. A B C D

11. A B C D

12. A B C D

13. A B C D

14. A B C D

15. A B C D

16. A B C D

17. A B C D

18. A B C D

19. A B C D

20. A B C D

Question 1 MCQs [20 Marks]

Question 1: In pandas, what function is used to display basic statistics of a data frame?

a) summary()

b) info()

c) describe()

d) stats()

Question 2: What is the purpose of stemming in text processing?

a) Identifying named entities

b) Reducing words to their root form

c) Removing stop words

d) Normalizing text

Question 3: Outliers in the dataset can be identified visually using which of the following visualization method?

a) Histograms

b) Box Plots

c) Pie Charts

d) All of the above

Question 4: Which of the following python function is used to remove duplicate values in a dataframe?

a) isduplicated()

b) drop_duplicates()

c) remove_ duplicates()

d) None of the above

Question 5: What should we do if we have variables with different ranges in our dataset?

a) Normalize the variable values

b) Remove the variables with different ranges

c) Fill the variable values with mean values

d) None of the above

Question 6: What is lemmatization in text processing?

a) Removing punctuation marks

b) Converting words to lowercase

c) Reducing words to their base or dictionary form

d) Removing numerical digits

Question 7: Which of the following is a popular feature extraction technique in NLP that describes the occurrence of each word within a document?

a) Lemmatization

b) Bag of Words

c) Stemming

d) None of the above

Question 8: Which of the following is not an application of computer vision?

a) Image Classification

b) Face Recognition

c) Drone-based crop monitoring

d) None of the above

Question 9: What is the primary purpose of applying edge detection filters in image processing?

a) Enhancing image brightness

b) Detecting and highlighting object boundaries

c) Removing noise from the image

d) Adjusting image contrast

Question 10: In machine learning, what does regression aim to predict?

a) Categories or classes

b) Group memberships

c) Continuous numerical values

d) None of the above

Question 11: You are analyzing a large dataset of text documents and want to find the most commonly used words in the corpus. Which of the following techniques would be the most suitable for this task?

a) Lemmatization

b) N-grams analysis

c) Punctuation removal

d) Stop-word removal

Question 12: A company is preparing to publish a research report with images containing sensitive information, including people's faces that need to be concealed for confidentiality purposes. Which image processing technique would be the most suitable for achieving this goal while maintaining the integrity of the image?

a) Blurring

b) Splitting Channels

c) Resizing and Rotation

d) Gray Scaling

Question 13: A photographer wants to create a mirror image of a landscape photograph to improve its composition. Which image processing technique would be most suitable for this purpose?

a) Flipping

b) Splitting Channels

c) Resizing

d) Gray Scaling

Question 14: Consider a dataset with the following columns: 'Customer ID,' 'Age,' 'Income,' and 'Loan Default Status' (with values 'Yes' or 'No'). It is a Classification problem. What could be an appropriate choice for the label (Y variable) in a machine learning problem based on this dataset?

a) Label: 'Customer ID'

b) Label: 'Loan Default Status'

c) Label: 'Income'

d) Label: 'Age'

Question 15: In a decision tree, what does a leaf node represent?

a) A decision point

b) A probability distribution

c) An intermediate step in the decision-making process

d) A final outcome or classification

Question 16: What does the process of outlier detection involve during EDA?

a) Identifying extreme or uncommon observations in the dataset

b) Highlighting the most frequent data points

c) Deleting data points with high variability

d) Ignoring data points that fall within the interquartile range

Question 17: What characterizes numerical data?

a) Descriptive labels or categories

b) Counts or measurements

c) Binary values

d) Ordinal rankings

Question 18: What is the purpose of a box plot in exploratory data analysis?

a) Showing data distribution shape

b) Displaying quartiles, median, and outliers

c) Representing probability density

d) Highlighting trends in time series data

Question 19: What practical application of Natural Language Processing is?

a) Spam Detection

b) Language Translation

c) Siri and Cortana

d) All of the above

Question 20: A surveillance system is being set up to monitor a high-security facility. To optimize the storage space and processing resources, the system needs to reduce the size of the captured images while preserving critical details which is held by RGB combinations. Which image processing technique would be most suitable for this purpose?

[✓ a) Resizing]

b) Gray Scaling

c) Flipping

d) Blurring

Question 2 Short questions [15 Marks]

Question 2a [5 Marks]: Based on the given machine learning use cases, determine if the problem is supervised or unsupervised.

1. A retailer wants to segment its customer base into different groups for targeted marketing based on their shopping habits and preferences.

   Unsupervised

2. A company wants to predict the price of a new product they are launching based on historical prices data of similar products.

   Supervised

3. In a retail dataset containing customer information such as 'Age', 'Income', 'Spending Score,' and 'Purchase History,' (as shown below) the company aims to group customers based on their purchasing behavior and spending patterns. Without any predefined labels, which type of machine learning approach would be best suited to uncover distinct customer segments for targeted marketing strategies and personalized recommendations?

Age	**Income**	**Spending Score**	**Purchase History**
35	72000	75	High
45	80000	80	High
25	48000	40	Medium

   Unsupervised

4. You have a dataset containing customer purchase history, but there's no information about whether the purchases were successful or not. Would you use supervised or unsupervised learning to analyze this data?

Question 2-b [5 Marks]: Based on the given machine learning use cases, determine if the problem is a classification or regression or clustering one.

1. A weather forecasting service wants to predict the daily temperature in degrees Celsius (°C) for the next seven days in Islamabad city.

   Regression

2. A music streaming service wants to predict the genre of a song based on its audio features, such as tempo, melody, and rhythm.

   Classification

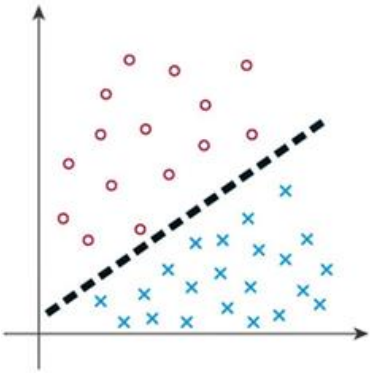
3. Grouping customers based on their purchasing behavior and demographics.

   Clustering

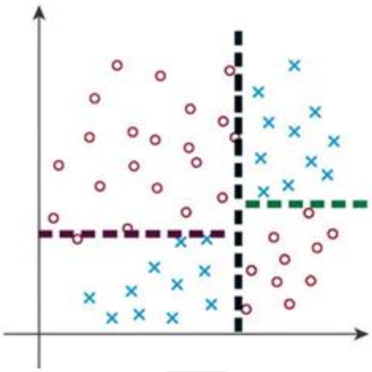
4. You are given a dataset with customer reviews, and your goal is to classify the sentiment of each review as positive, negative, or neutral. Would you choose regression, classification, or clustering for this task, and why?

   Classification

Question 2-c [2 Marks]: Consider the diagrams A and B below, what is the type (Regression or Classification) of both A and B.



A



B

A- Classification
B- Classification

Question-2d [3 points]:

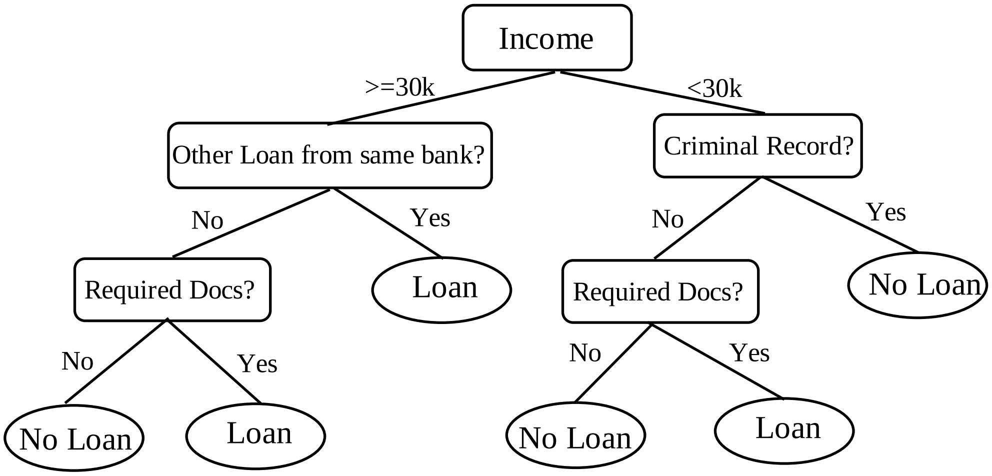
(1) List any two attribute selection measures that the decision tree model uses to select an attribute for the node split [2 points]

Information Gain

Gini Index

Gain Ratio

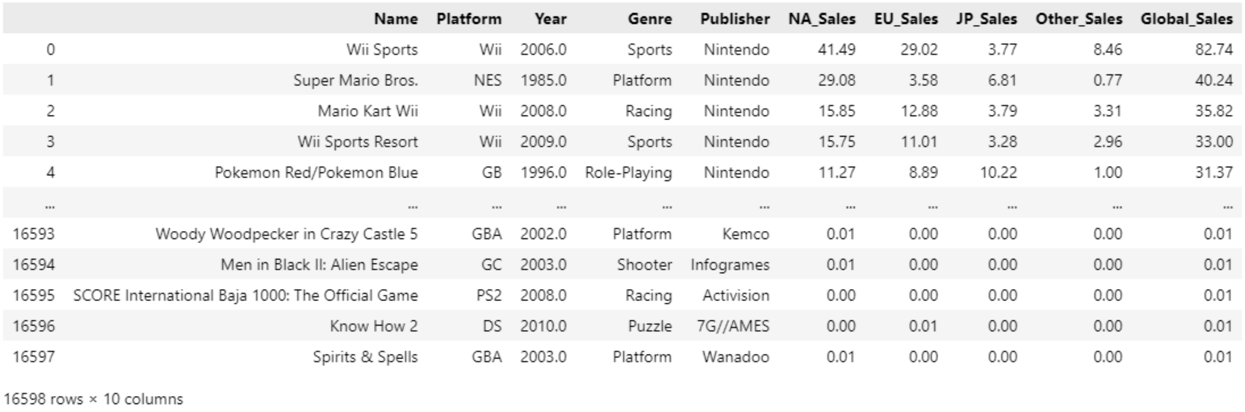
(2) Consider the following decision tree, write the class/classes name mentioned [1 point].



Loan
No Loan

Question 3 Exploratory Data Analysis [10 Marks]

Consider the following overview of a dataset containing video games that have sold more than 100,000 copies between 1980 and 2020. Column descriptions is also provided below



	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
...	...	...	...	...	...	...	...	...	...	...
16593	Woody Woodpecker in Crazy Castle 5	GBA	2002.0	Platform	Kemco	0.01	0.00	0.00	0.00	0.01
16594	Men in Black II: Alien Escape	GC	2003.0	Shooter	Infogrames	0.01	0.00	0.00	0.00	0.01
16595	SCORE International Baja 1000: The Official Game	PS2	2008.0	Racing	Activision	0.00	0.00	0.00	0.00	0.01
16596	Know How 2	DS	2010.0	Puzzle	7G//AMES	0.00	0.01	0.00	0.00	0.01
16597	Spirits & Spells	GBA	2003.0	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01

16598 rows × 10 columns

- Name - The games name
- Platform - Platform of the games release (i.e. PC,PS4, etc.)
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales

Answer the below questions.

Question 3-a [2 points]: Write the Python statements that will list down the sum of missing values for each column.

```
df.isnull().sum()
```

Question 3-b [3 points]: After identifying the missing values count, you analyze that the column Publisher has 20% missing values and Other_Sales have 10% missing values. Write python statements to drop all rows that have missing Publisher values and fill the missing values in Other_sales with the mean value of Other_sales.

```
m=df['Other_sales'].mean()
df['Other_sales'].fillna(m,inplace=True)
df['Publisher'].dropna(inplace=True)
```

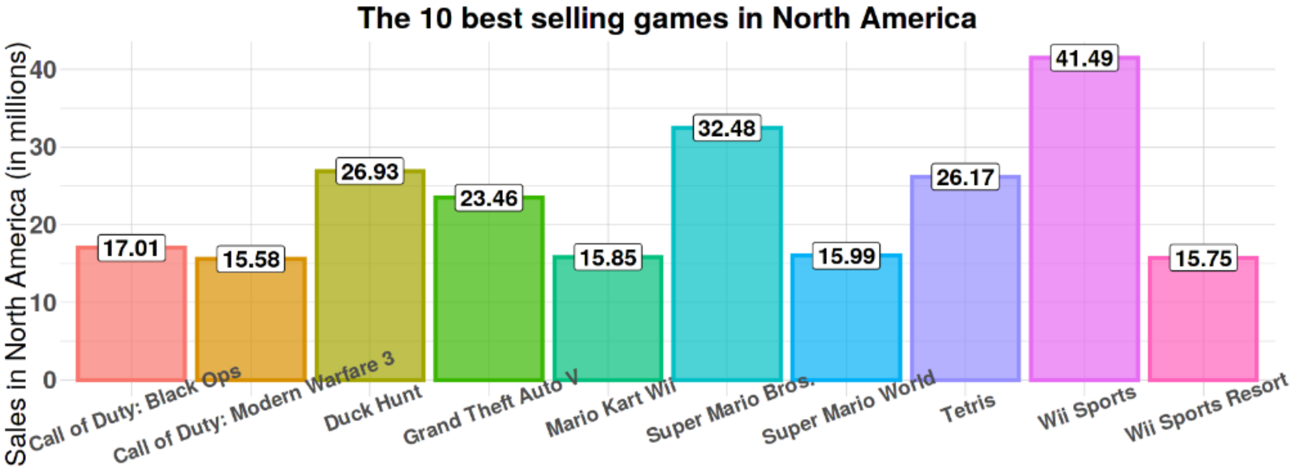
Question 3-c [1.5 points]: Write any 3 variables in your dataset that have categorical values.

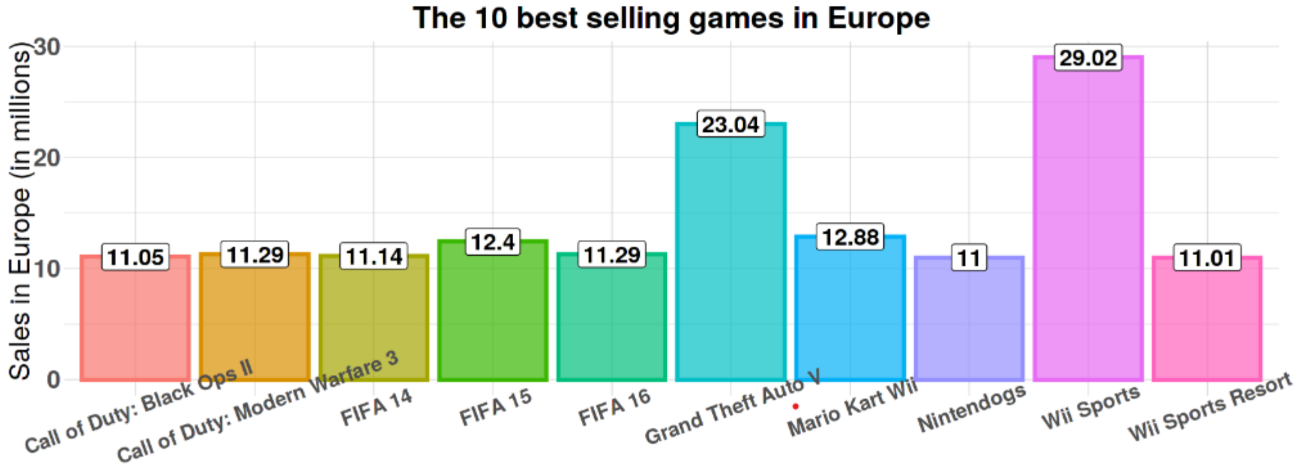
Platform, Genre, Publisher

Question 3-d [1 point]: Write python statements to get the descriptive statistics and summarized information of the dataset.

df.describe(), df.info()

Question 3-e [2.5 points]: Suppose an analyst generates the following charts of 10 best-selling games in North America and in Europe (name of game on x-axis). What inference can be drawn from the following figure?





Wii Sport is the most popular game in North America and is most popular in Europe as well.
Fifa games make the top 10 in Europe where the sports is more popular while it is not in top 10 games in North America.

Question 4 Performance Metrics [10 Marks]

Suppose you are working as a medical analyst at Control for Disease Control and Prevention (CDC) who wants to predict how many people are infected with a contagious virus (Covid-19) before they show symptoms and isolate them from the healthy population. The two classes for our target variable would be Sick (Covid positive) and Not Sick (Covid negative). You first tried the Decision Tree model and are now in the process of evaluating your model. Answer the below questions:

Question 4-a [2 points]: Assume you have 500 samples for your data analysis. From the dataset you know that the actual number of samples containing Covid positives is 160 and the actual number of samples without Covid is 340. In addition, sick people correctly predicted as sick by the model are 105, healthy people incorrectly predicted as sick by the model are 35, sick people incorrectly predicted as not sick by the model are 55 and healthy people correctly predicted as not sick by the model are 300. Construct the following confusion matrix.

	Actual Positive	**Actual Negative**
Predicted Positive	**105**	**35**
Predicted Negative	**55**	**300**

Question 4-b [1 point]: What is TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative) in your scenario.

TP	**105**
FP	**35**
TN	**300**
FN	**55**

Question 4-c [4 points]: Based on the confusion matrix calculated above, compute the following evaluation metrics (Show your working).

Accuracy:

(TP+TN)/(TP+FP+TN+FN) = (105+300)/500 = 405/500 = 0.81 or 81%

F1:

Precision = TP/(TP+FP) = 105/(105+35) = 105/140= 0.75 or 75%
Recall = TP/(TP+FN) = 105/(105+55) = 105/160= 0.656 or 65.6%

F1= 2 * Precision* Recall /(Precision + Recall) = 2*75*65.6/(75+65.6)= 69.9%

Question 4-d [1 point]: Is your dataset balanced (Just write Yes or No)?

No

Question 4-e [2 points]: Based on the evaluation metrics, what can you say about the model if it is good or bad? Suggest 1 strategy that you can utilize to increase the performance of your decision tree model.

Although Accuracy is good but as dataset is not balanced, accuracy is not a good measure to evaluate performance. F1 score tells us that model is not good.
To increase performance, change parameters of training.