COMSATS University Islamabad, Abbottabad Campus

Department of Computer Science

Mid Term Exam - Fall 2024

Class: BSE-7A

Date: 16/11/2024

Subject: CSC 461, Introduction to Data Science

Instructor: Dr. Ghulam Mujtaba

Total Time Allowed: 80 Mins

Max Marks: 50

Q1a: Please write the type of data each value against each attribute value/s: (8marks)

Data Record	Write: numeric, nominal, ordinal
a. 175 cm	
b. "Red," "Green," "Blue"	
c. "High," "Medium," "Low"	
d. 2023	
e. "Male," "Female"	
f. "First," "Second," "Third"	
g. 3.14159	
h. "Apple," "Banana," "Orange"	

Q1 b): For the following confusion matrix of a two class model,

	Predicted Positive		Predicted Negative	
Actual Positive	TP	80	FN	20
Actual Negative	FP	15	TN	85

Please fill in the following values after computing: (10 marks)

Accuracy	Error Rate	True Positives	True Negatives	False Positives	False Negatives	FPR	FNR	Precision	Recall

Q2 a). Given three nodes with the following class distributions: (6 marks)

Node 1: 40 instances of class A, 60 instances of class B

Node 2: 25 instances of class A, 75 instances of class B

Node 3: 90 instances of class A, 10 instances of class B

Calculate the Gini impurity for each node.

Q2 b): In a Decision Tree classification (12 marks)

i) briefly define overfitting.

ii) Mention two causes of overfitting.

iii) and how it can be reduced.

Q3a: In a Rule based Classification, briefly define the rule properties: (8 marks)

i) Mutually exclusive Rules
ii) Exhaustive Rules

Q3b: Consider the dataset: (6 marks)

	Age	Income
A	25	30000
B	30	45000
C	35	50000

We have a new data point, D, with the following attributes:

Age: 28, Income: 35000 → A

Determine its class using the kNN algorithm, with k = 1, using Euclidean Distance?