National University of Computer & Emerging Sciences, Karachi
Midterm One (Fall-2023)
28th Sept 2023, 10:00 AM – 11:00 AM

Course Code: AI-3002	**Course Name: Machine learning**
Instructor Name: Dr. Muhammad Farrukh Shahid, Sania Urooj	
Student Roll No:	**Section:**

Instructions:

- Return the question paper and make sure to keep it inside your answer sheet.
- Read each question completely before answering it. There are **3 questions and 3 pages**.
- In case of any ambiguity, you may make assumptions. But your assumption should not contradict any statement in the question paper.
- You are **not allowed to write** anything on the question paper (except your ID and group).

Time: 60 minutes. **Max Marks**: 50 Points

Question no 1 **CLO 1 [pt = 10]**

Identify whether the below statements are True or False. There will be a deduction by 0.5 for any wrong answer.

1. Training error is high when a machine learning model undergoes underfitting.
2. The training error or accuracy doesn't provide an (optimistically) biased estimate of the generalization performance in the machine learning model.
3. In Random Forest, each tree is trained independently on the entire dataset.
4. KNN algorithm is considered as a parametric model.
5. Hamming distance is just like the Manhattan distance applied to binary feature vectors.
6. Adaboost provides more weight to misclassified data points in each iteration.
7. Increasing the value of k in KNN generally results in a model with higher bias and lower variance
8. In KNN, the value of k represents the number of features in the dataset.
9. Gini index of 0.5 indicates maximum impurity in a binary classification problem.
10. If 50 Patients are actually diagnosed as diabetic, and the model correctly identifies 45 of them as diabetic. 50 patients are not diabetic, but the model incorrectly classifies 10 of them as diabetic. The F1-score of the model would be 85.6%.

Question no 2 **CLO 1 [pt= 10+ 10=20]**

a) Consider a Natural Language processing (NLP) application in which a chatbot communicates to the queries in a text form. Assume that there are two documents as shown in table.1 and the chatbot has to find the document class of Document 3 either alpha or beta while executing the KNN algorithm. Find the cosine similarity between the two documents.

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

Table.1 Words and Document

Words	Document 1	Document 2	Document 3
Machine	1	1	1
Learning	1	0	1
is	1	1	0
an	1	1	0
interesting	1	1	0
not	0	1	1
course	0	1	1
Classification class	**Alpha**	**Beta**	?

b) Use KNN algorithm with k = sqrt(9) to classify the signal. Use Euclidean distance, Manhattan distance and weight feature distances with weight 0.6. Compare the results obtained with each distance.

IoT Devices	Noise level	Signal strength	Class
1	34	50	Lower power device
2	45	46	Lower power device
3	43	34	High power device
4	32	28	Lower power device
5	34	45	??

Question no 3 **CLO 2 [pt = 10 + 10 = 20]**

a) You are tasked with building a decision tree model to predict the likelihood of infection in patients based on the given medical dataset. The dataset contains information about fever, cough, breathing issue and Class label (0 for not infected, 1 for infected). Choose the first decision node for the tree.

Fever	Cough	Breathing issue	Infected
No	No	No	0
Yes	Yes	Yes	1
Yes	Yes	No	0
Yes	No	Yes	1
Yes	Yes	Yes	1
No	Yes	No	0
Yes	No	Yes	1
Yes	Yes	No	0

b) In the financial industry, one crucial task is assessing an applicant's creditworthiness to decide whether to approve or deny a loan application. Credit scoring models help banks and lending institutions make informed decisions. Therefore, Implement Bagging by creating an ensemble of decision trees. Once base models are trained, each decision tree classifier makes its own prediction on unseen data as shown in the table below.
Out of multiple outputs produced by the classifiers C1, C2 and C3, the majority output is chosen to be the final result of the model.

Assume Any value < 0.5, belongs to 0 and >=0.5, belongs to 1

C1	C2	C3	Average	Product	Minimum	Maximum	Majority
0.2	0.5	0.7					
0.1	0.6	0.2					
0.9	0.4	0.1					
0.2	0.3	0.4					