Machine Learning (R1UC525B)

Assignment 2

Submission Deadline: 3/1/2026

Numericals:

Dataset 1: Given Dataset

Consider the following dataset for predicting whether a student will PASS or FAIL an exam.

Student	Attendance	Internal Marks	Assignment Submitted	Result
S1	High	High	Yes	Pass
S2	High	Low	Yes	Pass
S3	Low	High	Yes	Pass
S4	Low	Low	Yes	Fail
S5	High	Low	No	Fail
S6	Low	High	No	Fail
S7	High	High	No	Pass
S8	Low	Low	No	Fail

Q1.

1. Compute the entropy of the target variable (Result).
2. Calculate Information Gain for each attribute.
3. Select the root node based on maximum Information Gain.
4. Draw the decision tree.

Q2.

1. Compute the Gini Index for each attribute.
2. Select the best split based on minimum Gini Index.
3. Compare the root attribute obtained with that from Information Gain.
4. Comment on any differences observed.

Q3.

1. Grow the decision tree fully using either Information Gain or Gini Index.
2. Identify branches that may lead to overfitting.
3. Explain why fully grown decision trees tend to overfit.

Q4.

A validation dataset is provided below:

Validation Set:
V1: (High, High, Yes) → Pass
V2: (Low, Low, Yes) → Fail
V3: (High, Low, No) → Fail
V4: (Low, High, No) → Fail

Tasks:
1. Explain the concept of Reduced Error Pruning (REP).
2. Using the validation set, evaluate pruning of at least one internal node.
3. Compare classification accuracy before and after pruning.
4. Justify whether pruning should be accepted or rejected.

Q5.

Critically analyze decision tree learning with respect to:
1. Information Gain vs Gini Index
2. Overfitting in decision trees
3. Importance of pruning
4. Effectiveness of Reduced Error Pruning in real-world datasets

Dataset 2: Given Dataset

Point	Coordinates
P1	(2, 3)
P2	(3, 4)
P3	(3, 2)
P4	(4, 3)

P5	(10, 10)
P6	(11, 10)
P7	(10, 11)
P8	(25, 30)

Q1.

Apply K-means clustering with k = 2.
Initial centroids:
C1 = P1 (2,3)
C2 = P5 (10,10)

Tasks:
1. Compute Euclidean distance of each point from both centroids.
2. Perform two iterations of K-means.
3. Report cluster membership and updated centroid positions.
4. Comment on the influence of point P8 on centroid movement.

Q2.

Apply K-medoid clustering with k = 2.
Initial medoids: P2 and P6.

Tasks:
1. Assign points to nearest medoid.
2. Evaluate possible medoid swaps.
3. Identify final clusters and medoids.
4. Compare results with K-means.

Q3.

Perform Agglomerative Hierarchical Clustering using:
Single linkage and Euclidean distance.

Tasks:
1. Show step-by-step merging.
2. Draw or describe dendrogram.
3. Cut dendrogram to form two clusters.
4. Explain placement of point P8.

Q4.

Apply DBSCAN with:
ε = 1.8, MinPts = 3.

Tasks:
1. Identify core, border, and noise points.
2. Determine clusters.
3. Identify noise/outliers.
4. Explain DBSCAN behavior for point P8.

Q5.

Critically analyze all four clustering techniques with respect to:
1. Sensitivity to outliers
2. Cluster shape assumptions
3. Parameter requirements
4. Interpretability and real-world suitability

Conclude which algorithm is most appropriate and which should be avoided for this dataset.

Dataset 3: Given Dataset

Consider the following 2-D dataset and clustering result obtained using a clustering algorithm (k = 2).

Point	Coordinates	Cluster Label
P1	(1, 1)	C1
P2	(1, 2)	C1
P3	(2, 1)	C1
P4	(8, 8)	C2
P5	(9, 8)	C2
P6	(8, 9)	C2

Q1.

1. For point P1, compute:
   a) a(i): Average intra-cluster distance
   b) b(i): Average nearest inter-cluster distance

2. Calculate the Silhouette coefficient s(i) for P1.
3. Repeat the calculation for any one point from cluster C2.

Q2.

1. Compute the Silhouette score for all data points.
2. Calculate the average Silhouette score for the clustering.

Q3.

Based on the computed Silhouette values:
1. Interpret what values close to +1, 0, and −1 indicate.
2. Comment on the quality of clustering obtained in this dataset.

Q4.

Answer the following:
1. How does the Silhouette metric help in selecting the optimal number of clusters?
2. What are the limitations of Silhouette score for non-spherical clusters?
3. Compare Silhouette score with at least one other cluster validation metric.

Case 4: Consider a GA used to maximize the function:

$$f(x) = x^2$$

where $x$is represented using 5-bit binary encoding.

Chromosome	Binary String
C1	01001
C2	01110
C3	10100

1. Decode each chromosome into decimal value.
2. Compute the fitness value for each chromosome.

3. Identify the best chromosome of the current generation.

Case 5: Given the following population:

Chromosome	Fitness
C1	10
C2	20
C3	30
C4	40

a) Compute selection probability for each chromosome using roulette wheel selection.
b) Which chromosome has the highest chance of selection and why?

Case 6:

Two parent chromosomes are:

- Parent 1: 1100101
- Parent 2: 1011010

a) Perform single-point crossover at position 4.
b) Apply bit-flip mutation on the second offspring at position 6.
c) Write the final offspring chromosomes.

Descriptive

1. Explain the agent–environment interaction loop in Reinforcement Learning with an example.
2. What is meant by episodic and continuous tasks in RL?
3. Describe the exploration vs exploitation dilemma. Why is it important?
4. What is a Markov Decision Process (MDP)? List its components.

5. Why is Reinforcement Learning considered suitable for real-world problems like robotics and game playing?
6. Explain the role of the following parameters in Q-Learning:
7. Learning rate (α)
8. Discount factor (γ)
9. What is a Q-table? When does it become impractical to use?
10. How does Q-Learning update its knowledge during learning?
11. Explain the phases of genetic algorithm.

12. List the main components of an FFNN and briefly explain the role of each.
13. What is the purpose of weights and biases in an FFNN?
14. Explain the function of the input layer**,** hidden layer**,** and output layer**.**