



CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

University Examination November 2024  
B.Tech (CS) - V

MACHINE LEARNING [CSE303]

Marks: 70

Duration: 195 mins.

Section - I

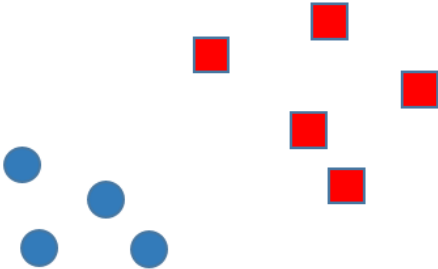
Answer all the questions.

Section Duration: 40 mins

1

How many number of neuron(s) is/are required to solve following problem?

(1)

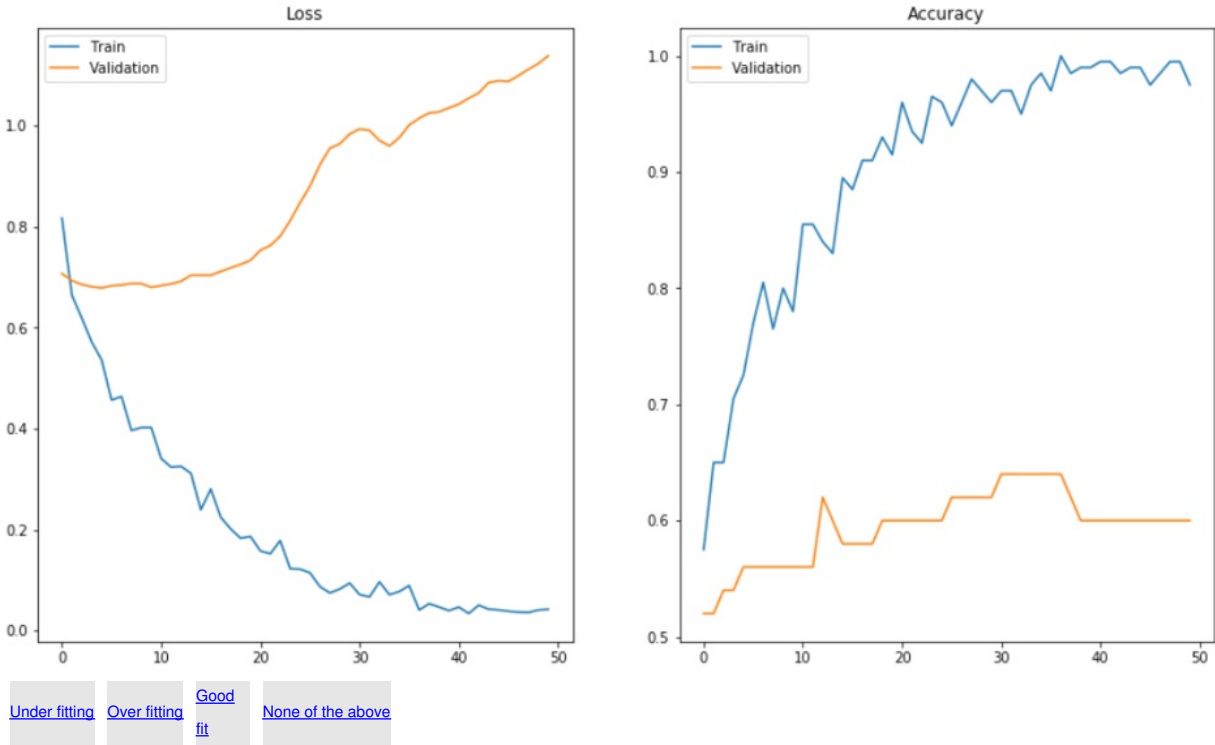


[2](#) [3](#) [4](#) [1](#)

2

A machine learning model is used to predict heart disease based on a dataset that includes features such as age, cholesterol levels, blood pressure, and lifestyle factors. After training a model, following loss and accuracy curves are observed. Which situation is occurred?

(1)



[Under fitting](#) [Over fitting](#) [Good fit](#) [None of the above](#)

3

A Health company wants to develop a machine learning model to predict whether patients will be diagnosed with diabetes based on historical health data. The data consists of labeled examples, with each patient's age, BMI, blood pressure, and glucose levels, along with a label indicating whether they were diagnosed with diabetes (1) or not (0). Based on this scenario, which of the following is correct?

(1)

[Use a clustering algorithm like K-Means to group patients based on their health data and predict the diagnosis](#)

[Use a Logistic Regression or Decision Trees, as the task involves predicting a labeled outcome](#)

[Use an unsupervised learning approach like Principal Component Analysis \(PCA\) to directly predict whether a patient has diabetes](#)

[Use reinforcement learning to teach the model to maximize its reward by correctly predicting diabetes diagnoses over time](#)

4

A machine learning model is built to predict customer churn for a subscription-based service. Which of the following is the correct sequence of steps to design and build the model?

(1)

[Collect data, Split data into training and testing sets, Train model, Define problem, Evaluate model, Select features, Deploy model](#)

[Define problem, Collect data, Preprocess data, Split data into training and testing sets, Train model, Evaluate model, Deploy model](#)

[Preprocess data, Collect data, Define problem, Train model, Split data into training and testing sets, Evaluate model, Deploy model](#)

[Define problem, Split data into training and testing sets, Train model, Evaluate model, Collect data, Deploy model](#)

5

A researcher trained four machine learning models using 1 million samples. Researcher noticed that one of the algorithms took longer testing time compared to the others. Which algorithm is most likely to have the highest testing time?

(1)

[Decision Tree](#)

[Logistic Regression](#)

[Support Vector Machine \(SVM\)](#)

[K-Nearest Neighbors \(KNN\)](#)

6

Which of the following best describes the goal of reinforcement learning?

(1)

[Minimize the loss function](#)

[Maximize the immediate reward at each step](#)

[Maximize the cumulative reward over time](#)

[Minimize the distance between the prediction and the actual output](#)

7

In the context of reinforcement learning, what is an "exploration vs exploitation" dilemma?

(1)

[Choosing between multiple actions based on maximum immediate reward](#)

[Balancing between trying new actions to discover better rewards or exploiting known actions that give the highest current reward](#)

[Optimizing a deep learning network versus a shallow network](#)

[Choosing between batch learning and online learning](#)

8

A car manufacturing company is working on an autonomous driving system. The system needs to navigate through various traffic conditions, including stop signs, pedestrians crossing, and traffic lights. The company wants the system to learn how to drive safely by receiving feedback for its actions in different driving scenarios. Which machine learning approach would be most suitable for training the autonomous driving system?

(1)

[Supervised Learning](#)

[Reinforcement Learning](#)

[Unsupervised Learning](#)

[Semi-Supervised Learning](#)

9

A research organization wants to conduct a pre-election poll by interviewing voters across the country. They divide the country into regions and select every 10th voter from a list of registered voters in each region. Which sampling method is being applied in this case?

(1)

[Systematic Sampling](#)

[Simple Random](#)

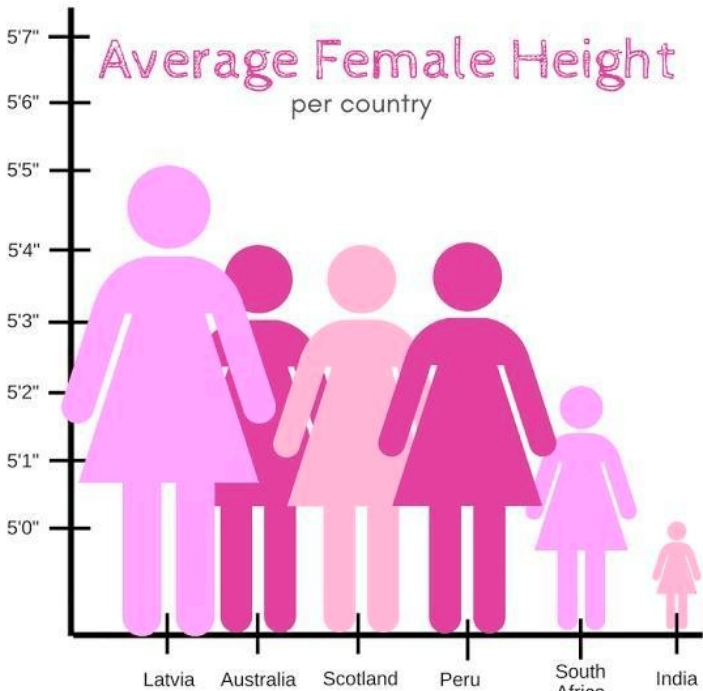
[Quota Sampling](#)

[Cluster Sampling](#)

10

Based on the information in the given picture, what is misleading about this graph ?

(1)

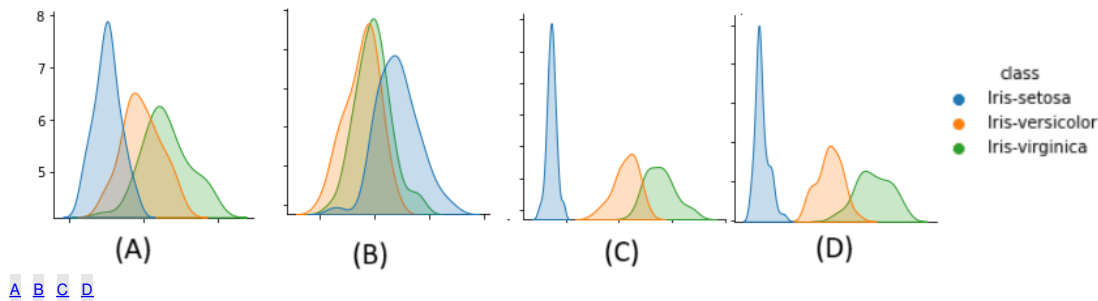


- [The scale of height is inconsistent.](#)
- [The x-axis should represent the numerical data.](#)
- [The graph does not indicate the average age of the women.](#)
- [The y-axis has not started from 0](#)

11

The Iris dataset contains three classes and four features named A, B, C, and D. The following graphs display the distributions of feature values for each class on x-axis and feature values on y-axis. Based on the observations from these graphs, which feature could be excluded?

(1)



[A](#) [B](#) [C](#) [D](#)

12

A company is analyzing the ages of customers who recently signed up for their loyalty program. The ages are recorded as follows: 1,1,2,3,5,7,7,8,10,12,15.

(1)

What is the IQR value?

[2](#) [10](#) [7](#) [8](#)

13

A marketing team at a retail company wants to segment their customer base to tailor their advertising strategies more effectively. They decide to use clustering techniques on customer data, which includes features such as age, income, shopping frequency, and product preferences. After running a clustering algorithm, they find distinct groups of customers with similar characteristics. What is the primary benefit of using clustering techniques in this scenario?

(1)

- [It provides a single solution that fits all customers.](#)
- [It eliminates the need for any marketing strategies.](#)
- [It guarantees increased sales for all customer segments.](#)
- [It helps identify underlying patterns in customer behavior.](#)

14

A data scientist working for a healthcare organization is tasked with predicting the likelihood of patients developing a particular disease. The dataset consists of hundreds of features, including age, gender, lifestyle factors, and various medical test results. To reduce the complexity of the model and avoid overfitting, the data scientist decides to apply Principal Component Analysis (PCA) before training the model. Why might PCA be beneficial in this scenario?

(1)

- [PCA increases the number of features to improve accuracy.](#)
- [PCA removes highly correlated features that add significant variance.](#)
- [PCA reduces the dimensionality of the dataset while preserving as much variance as possible.](#)
- [PCA enhances the interpretability of the individual features](#)

15

A company wants to classify printed letters - A, B, and C. They have collected 100 samples of each letter. The classification model needs to be deployed on a resource constrained device. Considering these situation, which type of learning approach would be most suitable for this task?

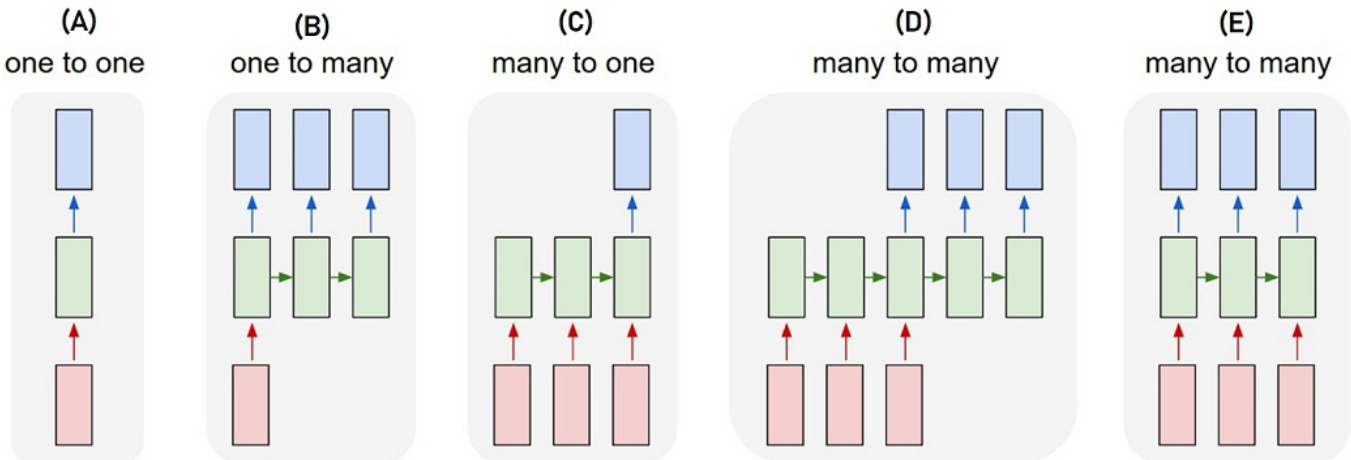
(1)

- [Machine Learning classification algorithms only](#)
- [Deep Learning CNN algorithm](#)
- [Machine Learning classification algorithms with feature extraction techniques combined with a classification algorithm](#)
- [Deep Learning algorithms with feature extraction techniques](#)

16

Which network structure would best fit for sentiment analysis and machine translation task from the following figure ?

(1)



- [Sentiment analysis: A and Machine translation: B](#)
- [Sentiment analysis: C and Machine translation: D](#)
- [Sentiment analysis: B and Machine translation: E](#)
- [Sentiment analysis: D and Machine translation: A](#)

17

Consider a simple neural network with a single neuron. The input to the neuron is  $x=0.5$ , the target output is  $t=1.0$ , the initial weight  $w=-0.4$ . and the bias term  $b=0.1$ . The neuron uses a relu activation function. The learning rate  $\alpha=0.05$ . What will be the output from the neuron?

(2)

[-0.1](#) [0](#) [0.1](#) [0.3](#)

18

Given the following dataset and initial centroids, determine the cluster assignments for points C and E after the first iteration of the K-means clustering algorithm: (2)

Dataset: A(1, 2), B(2, 3), C(3, 1), D(8, 8), E(9, 9)

Initial Centroids: C1(3, 2) and C2(8.5, 8.5)

C belongs to C1 cluster and E belongs to C2 cluster

C belongs to C2 cluster and E belongs to C1 cluster

Both points belong to C1 cluster

Both points belong to C2 cluster

Section - II

Answer all the questions.

1

Given the dataset containing physical measurements of abalones, the task is to predict the age of abalone using machine learning techniques. (5)

The dataset includes attributes such as Gender, length, diameter, height, and several weight measurements. The number of rings is the primary predictor for determining the age of the abalone, where adding 1.5 to the number of rings gives the actual age in years.

Sr. No.	Gender	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	No. of Rings
1	F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
2	F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	8
3	M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9
4	M	0.45	0.32	0.1	0.381	0.1705	0.075	0.115	2
5	F	0.55	0.415	0.135	0.7635	0.318	0.21	0.2	1
6	M	0.665	0.525	0.165	1.338	0.5515	0.3575	0.35	6
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
4082	M	0.21	0.15	0.05	0.042	0.0175	0.0125	0.015	4

Using this dataset, answer the following questions:

- What should be the size of training and testing set?
- Which is the target column? Write name of it.
- How many independent variables are in the given dataset? Write name of them.
- Which type of machine learning can be applied to solve this problem? What kind of task can be performed for the given data?
- Does categorical variable exist in the given data? If yes, mention it.

2

There is a dataset of housing sales in a particular region. The dataset contains the following columns for six recent sales: (5)

House ID	House Name	Location	Area (sq ft)	Bedrooms	Bathrooms	Sales Price (\$)
1	Shanti Nivas	Central	2000	3	2	300,000
2	Anand Bhavan	Suburban	3500	5	4	600,000
3		Central	800	2	a	90,000
4	Sukh Sagar	Downtown	1500	3	2	1,500,000
5	Lakshmi Kutir	Suburban	2500	4	3	350,000
6	Visamo	Downtown	1600			195000

- Analyse the dataset and identify any issues related to data pre-processing.
- For each identified issue in the dataset, explain how you would handle it.
- Perform an analysis of the 'Sales Price' column. Identify any potential outliers using an appropriate method. Explain how you would handle these outliers and provide justification for whether you would remove or retain them based on the nature of the data.

3

A bank has detected a new type of fraud over the last 48 hours, during which 120,000 transactions were processed. After manual verification, it was found that 14 transactions were fraudulent. To automate the detection of such fraudulent transactions in the future, the bank plans to deploy a fraud-detection system. (5)

Two competing systems, System A and System B, have been proposed for this task. Both systems use different classification algorithms. The performance of each system is evaluated based on the 120,000 transactions, and the bank has provided the following results:

System A Performance:

	Predicted Fraud	Predicted Not Fraud
Actual Fraud	5	9
Actual Not Fraud	1	119,985

System B Performance:

	Predicted Fraud	Predicted Not Fraud
Actual Fraud	12	2
Actual Not Fraud	5	119,971

- Compare the performance of System A and System B based on the evaluation matrices of above systems.
- Which system would be preferable by bank and why?

4

An organization is analyzing the impact of its promotional budget on sales performance to optimize future marketing strategies. The company has collected the following data: (5)

Promotion Amount Spent in Thousands	Sales in Thousands
1	2
2	3
3	5
4	4
5	6
6	7

Plot the data on an appropriate chart to observe the relationship between promotion amount and sales. Which relationship is observed between variables?

Build Machine Learning model to predict the sales if the organization spends 8,000 on promotion.

OR

5

(5)

A zoo wants to classify different types of Species based on their attributes. The zoo has collected data on various species, including their ability such as swim, fly, crawl, and its class. Apply the Naive Bayes algorithm to determine the class of the following new sample.

Page #3

New Sample:  
· **Swim**: Slow  
· **Fly**: Rarely  
· **Crawl**: No

Sr. No	Swim	Fly	Crawl	Class
1	Fast	No	No	Fish
2	Fast	No	Yes	Animal
3	Slow	No	No	Animal
4	Fast	No	No	Animal
5	No	Short	No	Bird
6	No	Short	No	Bird
7	No	Rarely	No	Animal
8	Slow	No	Yes	Animal
9	Slow	No	No	Fish
10	Slow	No	Yes	Fish
11	No	Long	No	Bird
12	Fast	No	No	Bird

6

A medical research team is analyzing a dataset to study the relationship between blood group, the presence of fever, and blood test results. The blood test results are categorized as either positive or negative. The goal is to determine which feature most significantly influences the blood test outcome. Using the contingency tables provided below, apply an appropriate feature selection method **to identify the most relevant feature**. Consider Significance Level ( $\alpha$ ) is 0.01 and  $df = 1$ , the critical value is 6.635.

(5)

Blood Group (Feature 1)	Positive Test (1)	Negative Test (0)
AB	70	30
O	100	60

Fever (Feature 2)	Positive Test (1)	Negative Test (0)
Yes	120	10
No	20	50

OR

7

A data science team is working on a predictive model using a dataset with two features (X1 and X2) and one target variable (Y). They aim to determine the feature selection method to evaluate the relationship between the features and the target. Describe the nature of the variables in this dataset. Suggest the feature selection method. Calculate the relationships between the features and the target variable in this dataset using the chosen feature selection method. Which feature would be the most appropriate for predictive model?

(5)

Sample	X1 (Feature 1)	X2 (Feature 2)	Y (Target)
1	3	4	3
2	2	7	4
3	4	5	5
4	8	6	6
5	12	5	8
6	10	4	9

Section - III

Answer all the questions.

1

A large healthcare company wants to use machine learning and AI to optimize patient care and improve sustainability. The company has collected vast amounts of data on patient health records, medical imaging, disease patterns, and treatment outcomes. They are considering using a combination of Natural Language Processing (NLP) to process medical reports, Deep Learning algorithms (such as CNNs and RNNs) to analyze image and time-series data from patient monitoring systems. Based on this case study, answer the following:

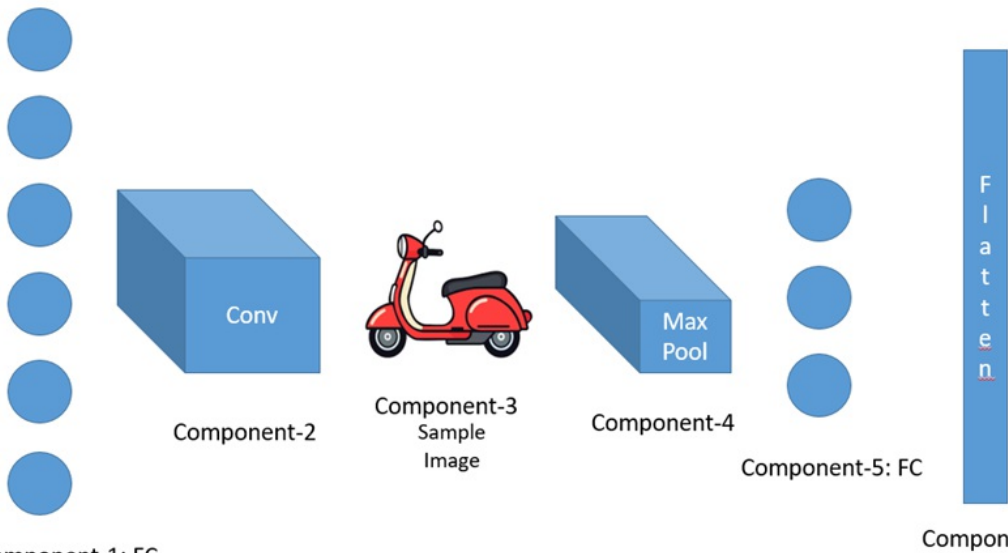
(5)

- How can CNNs and RNNs be used in this scenario to enhance healthcare practices?
- Explain the role of NLP in extracting valuable insights from medical reports and documents.
- Do you think that Explainable AI and Responsible AI are important in this context? Why?

2

"AutoVision" is a company specializing in the development of intelligent transportation systems. They aim to build a deep learning model using Convolutional Neural Networks (CNN) to classify vehicle images into categories: Scooter, Auto, and Tractor. The team has collected a dataset of labeled vehicle images and is now in the process of designing the CNN model architecture. As part of the design process, AutoVision has identified the following components that will be used in the CNN model:

(5)



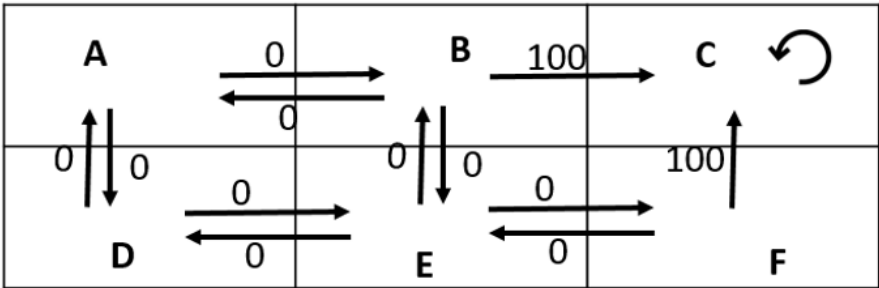
Component-1: FC

Create a CNN model by arranging these components in the correct sequence to ensure the functions for vehicle image classification. Justify the role of each in the overall architecture. Suggest the hyper-parameters. Also, clearly show that which components are used for feature extraction and which are for classification?

3

A robot is positioned in a 2x3 grid environment, as shown in the diagram. The grid consists of six states labeled A, B, C, D, E, and F. The robot can move in four directions: up, down, left, and right, but it cannot move outside the grid boundaries. The arrows in the diagram indicate possible transitions between states, with specific rewards associated with each transition. Consider  $\gamma = 0.8$ .

(5)



1. Construct a Reward Matrix.

2. Using the Q-learning algorithm, compute the Q-values for the following state transitions: From B to C; From A to B; and From B to A

3. After computing the Q-values for the specified transitions, show the updated Q-matrix reflecting these changes.

4

A retail store aims to analyze customer purchasing behavior to identify frequent item combinations that are often bought together. Find all frequent item-sets using the Apriori algorithm. The store has set a minimum support threshold of 2 and a minimum confidence threshold of 60% to uncover valuable purchasing patterns that can inform inventory management and marketing strategies.

(5)

The retail store also wants to analyze whether customers who purchase Bread (I1) and Milk (I3) together are also likely to purchase Sugar (I5). Provide recommendation on this analysis which will assist the store in designing effective product placement strategies and targeted promotions.

I1- Bread, I2- Butter, I3- Milk, I4-Cheese, I5-Sugar

TID	ITEMS
T1	I1, I3, I4
T2	I2, I3, I5
T3	I1, I2, I3, I5
T4	I2, I5
T5	I1, I3, I5

5

OR

A data analyst is tasked with exploring spatial relationships within a two-dimensional dataset to identify potential clusters using the DBSCAN algorithm. The dataset consists of five spatial points, and the parameters are set to  $Eps \leq 5$  and  $MinPts = 2$ .

(5)

Point	X	Y
P1	2	10
P2	2	5
P3	8	4
P4	5	8
P5	7	5

Which points are classified as **core points**, **border points**, and **noise points**?

6

Which clustering method should NOT be used for the following sample dataset? Justify.

(5)

Which are the types of hierarchical clustering algorithm? Explain any one. Which are the types of different linkage methods?



7

OR

In Support Vector Machine (SVM) algorithm, what is support vector? Draw a sample diagram to show support vectors. What is the importance of this algorithm compared to other classification algorithms? How following data can be solved using SVM? and Show resultant decision boundary.

(5)



-----End-----