CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

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

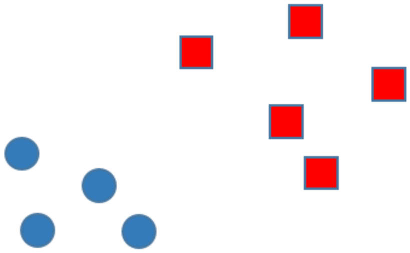
MACHINE LEARNING [CSE303]

Marks: 70 **Duration: 195 mins.**

Section - I

Answer all the questions. Section Duration: 40 mins

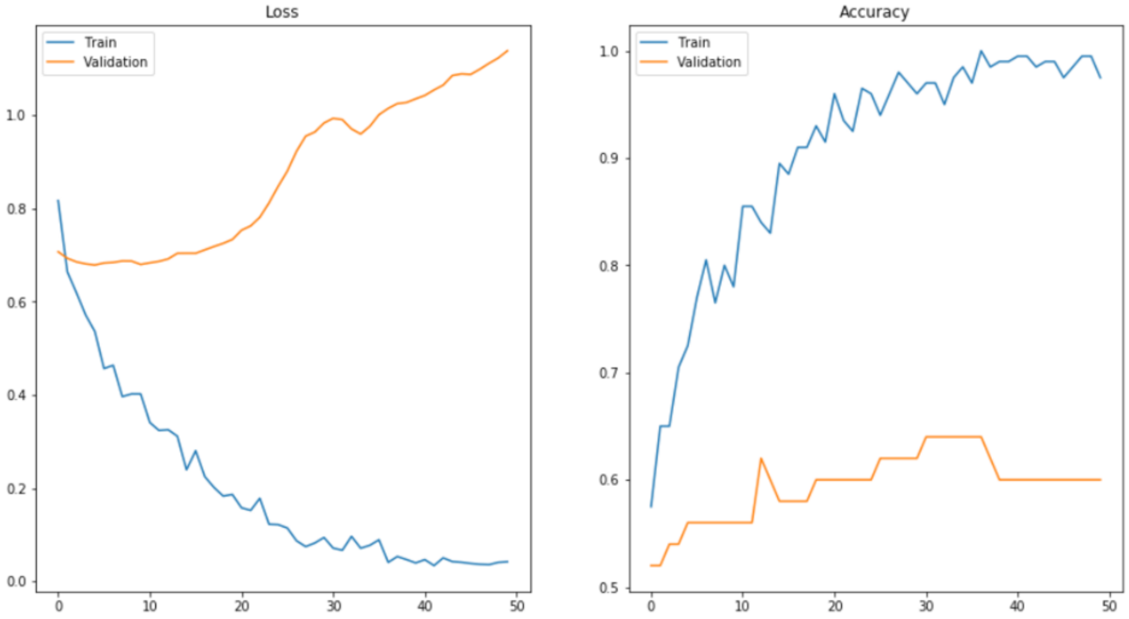
1 How many number of neuron(s) is/are required to solve followingproblem? (1)



1)	2		2)	3		3)	4		4)	1	

Course Outcome: "CO3" / Cognitive Level: "Analysis"

2 A machine learning model is used to predict heart disease based ona 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 isoccurred? (1)



1)	Under fitting		2)	Over fitting		3)	Good fit		4)	None of the above	

Course Outcome: "CO3" / Cognitive Level: "Analysis"

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

1)	Use a clustering algorithm like K-Means to group patients based on their health data and predict the diagnosis		2)	Use a Logistic Regression or Decision Trees, as the task involves predicting a labeled outcome		3)	Use an unsupervised learning approach like Principal Component Analysis (PCA) to directly predict whether a patient has diabetes		4)	Use reinforcement learning to teach the model to maximize its reward by correctly predicting diabetes diagnoses over time	

Course Outcome: "CO1" / Cognitive Level: "Application"

4 A machine learning model is built to predict customer churn for asubscription-based service. Which of the following is the correctsequence of steps to design and build the model? (1)

1)	Collect data, Split data into training and testing sets, Train model, Define problem, Evaluate model, Select features, Deploy model		2)	Define problem, Collect data, Preprocess data, Split data into training and testing sets, Train model, Evaluate model, Deploy model		3)	Preprocess data, Collect data, Define problem, Train model, Split data into training and testing sets, Evaluate model, Deploy model.		4)	Define problem, Split data into training and testing sets, Train model, Evaluate model, Collect data, Deploy model	

Course Outcome: "CO1" / Cognitive Level: "Remember"

5 A researcher trained four machine learning models using 1 millionsamples. Researcher noticed that one of the algorithms took longertesting time compared to the others. Which algorithm is most likelyto have the highest testing time? (1)

1)	Decision Tree		2)	Logistic Regression		3)	Support Vector Machine (SVM)		4)	K-Nearest Neighbors (KNN)	

Course Outcome: "CO3" / Cognitive Level: "Application"

6

Which of the following best describes the goal of reinforcementlearning?

1)	Minimize the loss function		2)	Maximize the immediate reward at each step		3)	Maximize the cumulative reward over time		4)	Minimize the distance between the prediction and the actual output	

(1)

Course Outcome: "CO5" / Cognitive Level: "Remember"

7

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

1)	Choosing between multiple actions based on maximum immediate reward		2)	Balancing between trying new actions to discover better rewards or exploiting known actions that give the highest current reward		3)	Optimizing a deep learning network versus a shallow network		4)	Choosing between batch learning and online learning	

(1)

Course Outcome: "CO5" / Cognitive Level: "Remember"

8

A car manufacturing company is working on an autonomous drivingsystem. The system needs to navigate through various trafficconditions, including stop signs, pedestrians crossing, and trafficlights. The company wants the system to learn how to drive safelyby receiving feedback for its actions in different drivingscenarios. Which machine learning approach would be most suitablefor training the autonomous driving system?

1)	Supervised Learning		2)	Reinforcement Learning		3)	Unsupervised Learning		4)	Semi-Supervised Learning	

(1)

Course Outcome: "CO5" / Cognitive Level: "Application"

9

A research organization wants to conduct apre-election poll by interviewing voters across the country. Theydivide the country into regions and select every 10th voter from alist of registered voters in each region. Which sampling method isbeing applied in this case?

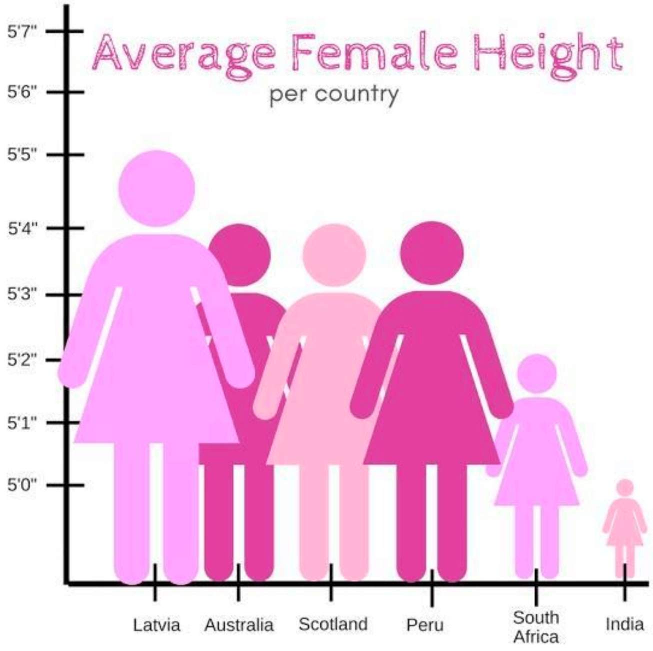
1)	Systematic Sampling		2)	Simple Random Sampling		3)	Quota Sampling		4)	Cluster Sampling	

(1)

Course Outcome: "CO2" / Cognitive Level: "Application"

10

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



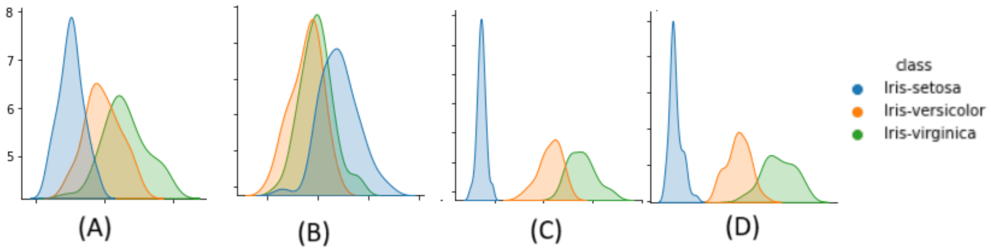
(1)

1)	The scale of height is inconsistent.		2)	The x-axis should represent the numerical data.		3)	The graph does not indicate the average age of the women.		4)	The y-axis has not started from 0	

Course Outcome: "CO2" / Cognitive Level: "Analysis"

11

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



(1)

1)	A		2)	B		3)	C		4)	D	

Course Outcome: "CO2" / Cognitive Level: "Evaluate"

12

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

(1)

What is the IQR value?

1)	2		2)	10		3)	7		4)	8	

Course Outcome: "CO2" / Cognitive Level: "Understand"

13

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

1)	It provides a single solution that fits all customers.		2)	It eliminates the need for any marketing strategies.		3)	It guarantees increased sales for all customer segments.		4)	It helps identify underlying patterns in customer behavior.	

(1)

Course Outcome: "CO4" / Cognitive Level: "Understand"

14

A data scientist working for a healthcareorganization is tasked with predicting the likelihood of patientsdeveloping a particular disease. The dataset consists of hundredsof features, including age, gender, lifestyle factors, and variousmedical test results. To reduce the complexity of the model andavoid overfitting, the data scientist decides to apply PrincipalComponent Analysis (PCA) before training the model. Why might PCAbe beneficial in this scenario?

1)	PCA increases the number of features to improve accuracy.		2)	PCA removes highly correlated features that add significant variance.		3)	PCA reduces the dimensionality of the dataset while preserving as much variance as possible.		4)	PCA enhances the interpretability of the individual features	

(1)

Course Outcome: "CO4" / Cognitive Level: "Application"

15

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

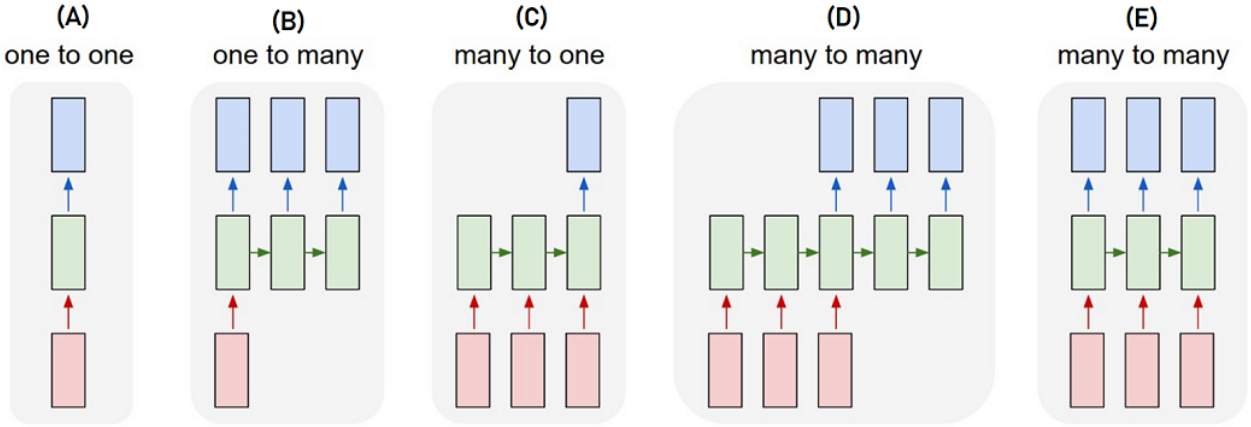
1)	Machine Learning classification algorithms only		2)	Deep Learning CNN algorithm		3)	Machine Learning classification algorithms with feature extraction techniques combined with a classification algorithm		4)	Deep Learning algorithms with feature extraction techniques	

(1)

Course Outcome: "CO5" / Cognitive Level: "Application"

16

Which network structure would best fit forsentiment analysis and machine translation task from the followingfigure?



(1)

1)	Sentiment analysis: A and Machine translation: B		2)	Sentiment analysis: C and Machine translation: D		3)	Sentiment analysis: B and Machine translation: E		4)	Sentiment analysis: D and Machine translation: A	

Course Outcome: "CO5" / Cognitive Level: "Application"

17

Consider a simple neural network with a single neuron. The input tothe neuron is x=0.5, the target output is t=1.0, the initial weightw= -0.4. and the bias term b=0.1. The neuron uses a relu activationfunction. The learning rate α=0.05. What will be the output fromthe neuron?

1)	-0.1		2)	0		3)	0.1		4)	0.3	

(2)

Course Outcome: "CO3" / Cognitive Level: "Understand"

18

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

1)	C belongs to C1 cluster and E belongs to C2 cluster		2)	C belongs to C2 cluster and E belongs to C1 cluster		3)	Both points belong to C1 cluster		4)	Both points belong to C2 cluster	

(2)

Course Outcome: "CO4" / Cognitive Level: "Understand"

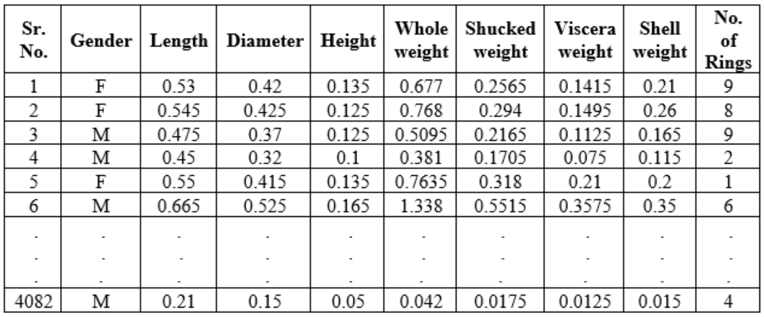
Section - II

Answer all the questions.

1

Given the dataset containing physical measurements of abalones, thetask is to predict the age of abalone using machine learningtechniques.
The dataset includes attributes such as Gender, length, diameter,height, and several weight measurements. The number of rings is theprimary predictor for determining the age of the abalone, whereadding 1.5 to the number of rings gives the actual age inyears.

(5)



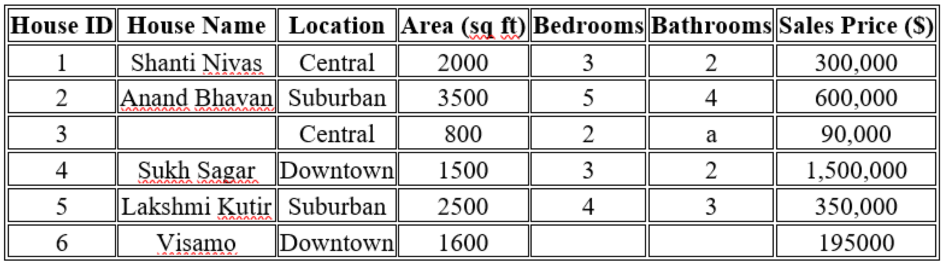
Sr. No.	Gender	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	No. of Rings
1	F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
2	F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	8
3	M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9
4	M	0.45	0.32	0.1	0.381	0.1705	0.075	0.115	2
5	F	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:

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

Course Outcome: "CO1" / Cognitive Level: "Analysis"

2 There is a dataset of housing sales in a particular region. Thedataset contains the following columns for six recent sales:



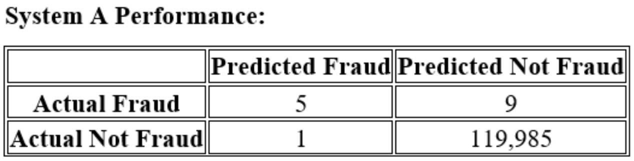
House ID	House Name	Location	Area (sq ft)	Bedrooms	Bathrooms	Sales Price ($)
1	Shanti Nivas	Central	2000	3	2	300,000
2	Anand Bhavan	Suburban	3500	5	4	600,000
3		Central	800	2	a	90,000
4	Sukh Sagar	Downtown	1500	3	2	1,500,000
5	Lakshmi Kutir	Suburban	2500	4	3	350,000
6	Visamo	Downtown	1600			195000

(5)

1. Analyse the dataset and identify any issues related to datapre-processing.
2. For each identified issue in the dataset, explain how you wouldhandle it.
3. Perform an analysis of the 'Sales Price' column. Identify anypotential outliers using an appropriate method. Explain how youwould handle these outliers and provide justification for whetheryou would remove or retain them based on the nature of thedata.

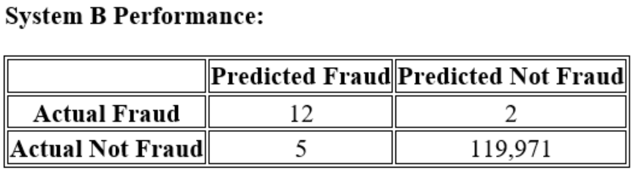
Course Outcome: "CO2" / Cognitive Level: "Analysis"

3 A bank has detected a new type of fraud over the last 48 hours,during which 120,000 transactions were processed. After manualverification, it was found that 14 transactions were fraudulent. Toautomate the detection of such fraudulent transactions in thefuture, the bank plans to deploy a fraud-detection system.
Two competing systems, System A and System B, have been proposedfor this task. Both systems use different classificationalgorithms. The performance of each system is evaluated based onthe 120,000 transactions, and the bank has provided the followingresults:



System A Performance:

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



System B Performance:

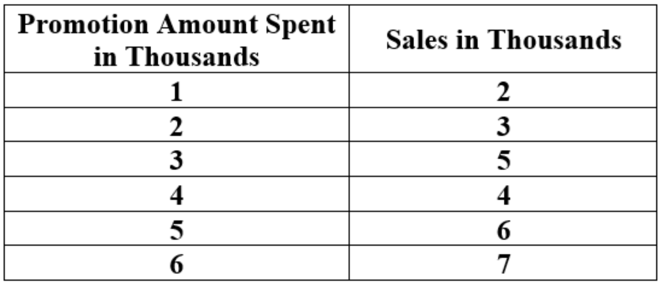
	Predicted Fraud	Predicted Not Fraud
Actual Fraud	12	2
Actual Not Fraud	5	119,971

(5)

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

Course Outcome: "CO3" / Cognitive Level: "Evaluate"

4 An organization is analyzing the impact of its promotional budgeton sales performance to optimize future marketing strategies. Thecompany has collected the following data: (5)



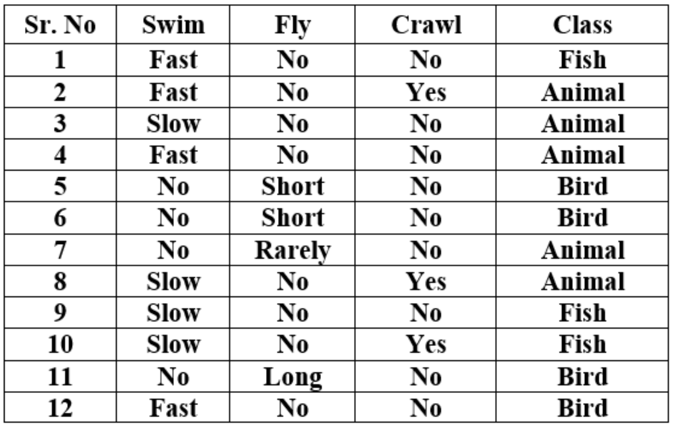
Promotion Amount Spent in Thousands	Sales in Thousands
1	2
2	3
3	5
4	4
5	6
6	7

Plot the data on an appropriate chart to observe the relationshipbetween promotion amount and sales. Which relationship is observedbetween variables?
Build Machine Learning model to predict the sales if theorganization spends 8,000 on promotion.

Course Outcome: "CO3" / Cognitive Level: "Application"

[OR]
5

A zoo wants to classify different types of Species based on theirattributes. 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 thefollowing new sample.
New Sample:
· **Swim**: Slow
· **Fly**: Rarely
· **Crawl:** No



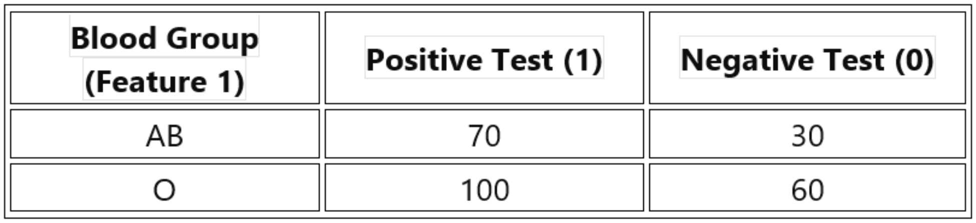
Sr. No	Swim	Fly	Crawl	Class
1	Fast	No	No	Fish
2	Fast	No	Yes	Animal
3	Slow	No	No	Animal
4	Fast	No	No	Animal
5	No	Short	No	Bird
6	No	Short	No	Bird
7	No	Rarely	No	Animal
8	Slow	No	Yes	Animal
9	Slow	No	No	Fish
10	Slow	No	Yes	Fish
11	No	Long	No	Bird
12	Fast	No	No	Bird

(5)

Course Outcome: "CO3" / Cognitive Level: "Application"

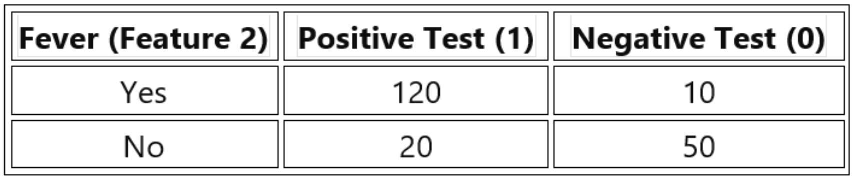
6

A medical research team is analyzing a dataset to study therelationship between blood group, the presence of fever, and bloodtest results. The blood test results are categorized as eitherpositive or negative. The goal is to determine which feature mostsignificantly influences the blood test outcome. Using thecontingency tables provided below, apply an appropriate featureselection method **to identify the most relevantfeature.**
Consider Significance Level (α) is 0.01 and df = 1, the criticalvalue is 6.635.



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

(5)



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

Course Outcome: "CO2" / Cognitive Level: "Application"

[OR]
7

(5)

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

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

Course Outcome: "CO2" / Cognitive Level: "Application"

Section - III

Answer all the questions.

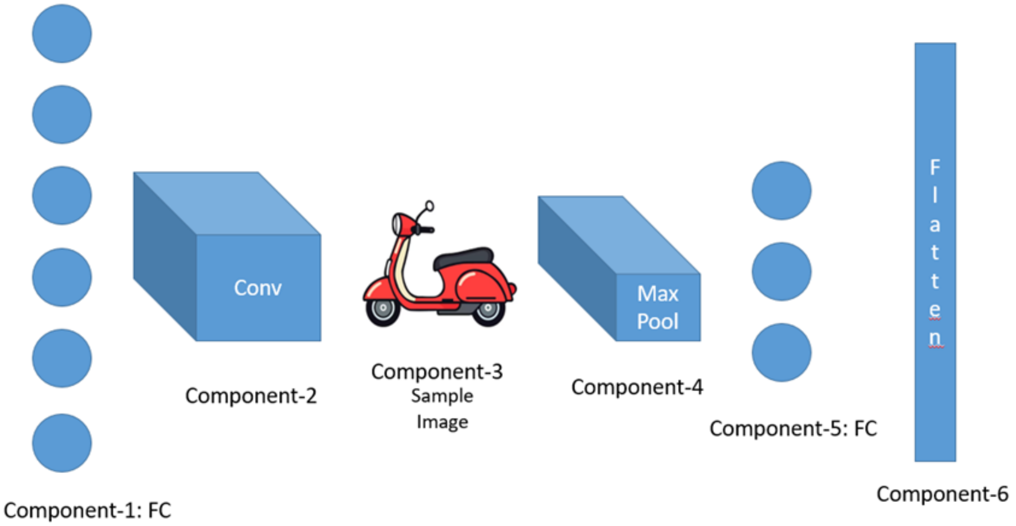
1 A large healthcare company wants to use machine learning and AI tooptimize patient care and improve sustainability. The company hascollected vast amounts of data on patient health records, medicalimaging, disease patterns, and treatment outcomes. They areconsidering using a combination of Natural Language Processing(NLP) to process medical reports, Deep Learning algorithms (such asCNNs and RNNs) to analyze image and time-series data from patientmonitoring systems.
Based on this case study, answer the following:

1. How can CNNs and RNNs be used in this scenario to enhancehealthcare practices?
2. Explain the role of NLP in extracting valuable insights frommedical reports and documents.
3. Do you think that Explainable AI and Responsible AI areimportant in this context? Why?

(5)

Course Outcome: "CO5" / Cognitive Level: "Understand"

2 "AutoVision" is a company specializing in the development ofintelligent transportation systems. They aim to build a deeplearning model using Convolutional Neural Networks (CNN) toclassify vehicle images into categories: Scooter, Auto, andTractor. The team has collected a dataset of labeled vehicle imagesand is now in the process of designing the CNN modelarchitecture.
As part of the design process, AutoVision has identified thefollowing components that will be used in the CNN model:

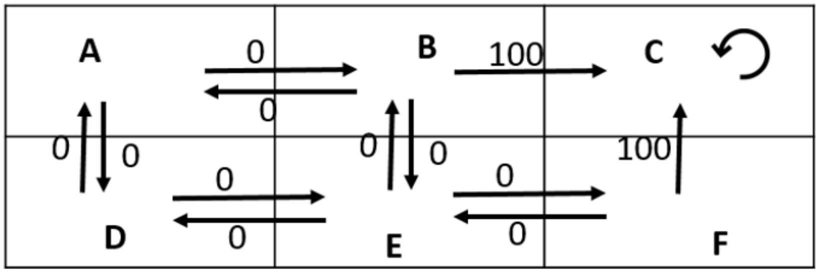


(5)

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

Course Outcome: "CO5" / Cognitive Level: "Create"

3 A robot is positioned in a 2x3 grid environment, as shown in thediagram. The grid consists of six states labeled A, B, C, D, E, andF. The robot can move in four directions: up, down, left, andright, but it cannot move outside the grid boundaries. The arrowsin the diagram indicate possible transitions between states, withspecific rewards associated with each transition. Consider γ =0.8.



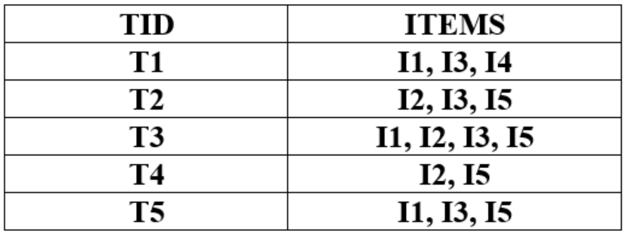
(5)

1. Construct a Reward Matrix.
2. Using the Q-learning algorithm, compute the Q-values for thefollowing state transitions: From B to C; From A to B; and From Bto A
3. After computing the Q-values for the specified transitions,show the updated Q-matrix reflecting these changes.

Course Outcome: "CO5" / Cognitive Level: "Application"

4 A retail store aims to analyze customer purchasing behavior toidentify frequent item combinations that are often bought together.Find all frequent item-sets using the Apriori algorithm. The storehas set a minimum support threshold of 2 and a minimum confidencethreshold of 60% to uncover valuable purchasing patterns that caninform inventory management and marketing strategies.
The retail store also wants to analyze whether customers whopurchase Bread (I1) and Milk (I3) together are also likely topurchase Sugar (I5). Provide recommendation on this analysis whichwill assist the store in designing effective product placementstrategies and targeted promotions. (5)

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

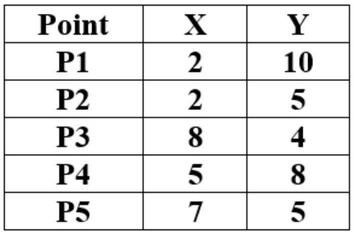


TID	ITEMS
T1	I1, I3, I4
T2	I2, I3, I5
T3	I1, I2, I3, I5
T4	I2, I5
T5	I1, I3, I5

Course Outcome: "CO4" / Cognitive Level: "Application"

[OR]
5

A data analyst is tasked with exploring spatial relationshipswithin a two-dimensional dataset to identify potential clustersusing the DBSCAN algorithm. The dataset consists of five spatialpoints, and the parameters are set to Eps ≤ 5 and MinPts = 2.



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

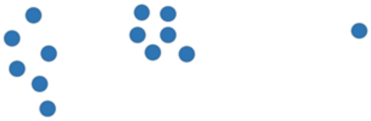
Which points are classified as **core points**,**border points**, and **noise points**?

(5)

Course Outcome: "CO4" / Cognitive Level: "Application"

6

Which clustering method should NOT be used for the following sampledataset? Justify.
Which are the types of hierarchical clustering algorithm? Explainany one. Which are the types of different linkage methods?

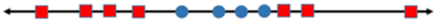


(5)

Course Outcome: "CO4" / Cognitive Level: "Understand"

[OR]
7

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



(5)

Course Outcome: "CO4" / Cognitive Level: "Understand"

-----End-----

Event Name	Subject
University Examination Nov -Dec-2024	CSE303:MACHINE LEARNING

#	Course Outcome	Description	Marks
1	CO1	Analyze and apply fundamental concepts of machine learning, including types, theory, and practices, to design effective learning systems.	7
2	CO2	Apply statistical data analysis techniques such as descriptive statistics, data cleaning, and visualization to pre-process and understand datasets for	19
3	CO3	Implement and evaluate supervised learning algorithms including classification and regression, using appropriate evaluation metrics such as confusion	20
4	CO4	Apply unsupervised learning methods such as clustering and dimensionality reduction to discover patterns and insights from unlabeled data.	24
5	CO5	Apply advanced topics in machine learning such as reinforcement learning, recent trends like natural language processing and deep learning, and their	20
		Total	**90**

* NA - CO Not Selected

Event Name	Subject
University Examination Nov -Dec-2024	CSE303:MACHINE LEARNING

#	Level	No of Questions	Marks (%)
1	Remember	3	3(3.33%)
2	Understand	7	21(23.33%)
3	Application	14	42(46.67%)
4	Analysis	5	13(14.44%)
5	Create	1	5(5.56%)
6	Evaluate	2	6(6.67%)
	Total	32	90 (100%)