Question Paper

Exam Date & Time: 05-Jan-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

SEVENTH SEMESTER B.TECH END SEMESTER MAKE UP EXAMINATIONS, JAN 2024

MACHINE LEARNING FOR DATA ANALYTICS [ICT 4056]

Marks: 50 **Duration: 180 mins.**

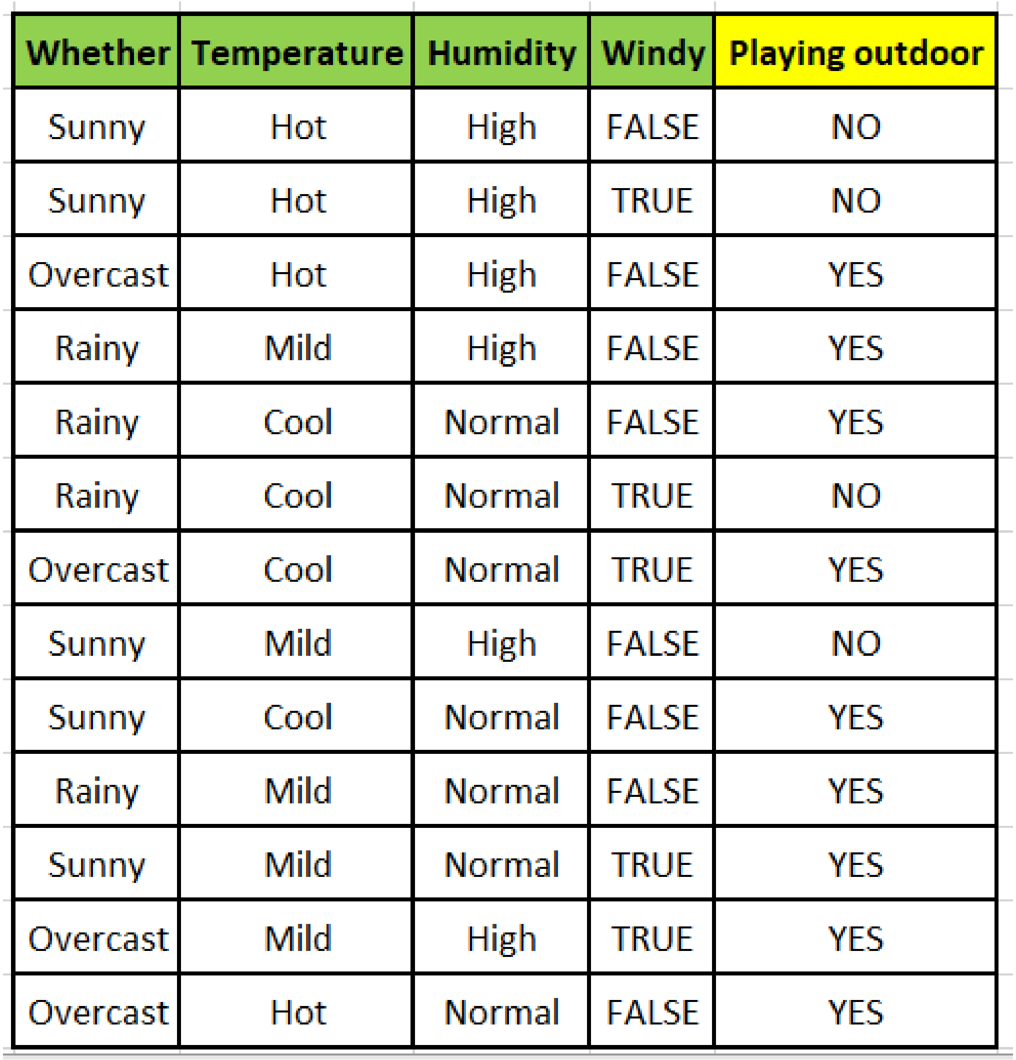
A

Answer all the questions.

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1) A) Based on the play condition data set as shown in Table 1, a decision has to be carried out by the organizers on conducting an outdoor game event. Draw the decision tree considering the feature "Temperature" as the root node and then with "Humidity". Compute the processing gain and Gini index. Also, draw the decision tree considering the "Humidity" feature as the root node and the "Temperature" feature as the second. (5)

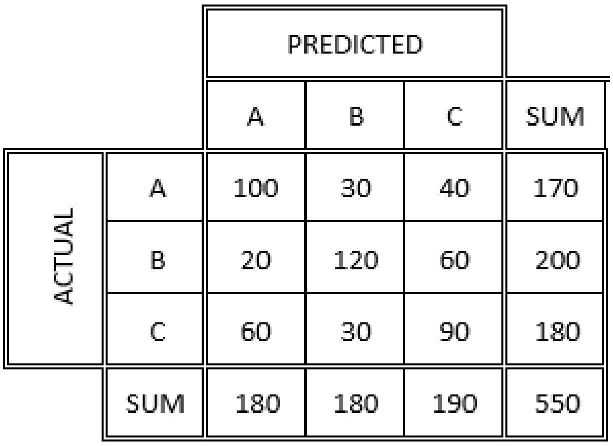
Table 1: Play Condition Data set



Whether	Temperature	Humidity	Windy	Playing outdoor
Sunny	Hot	High	FALSE	NO
Sunny	Hot	High	TRUE	NO
Overcast	Hot	High	FALSE	YES
Rainy	Mild	High	FALSE	YES
Rainy	Cool	Normal	FALSE	YES
Rainy	Cool	Normal	TRUE	NO
Overcast	Cool	Normal	TRUE	YES
Sunny	Mild	High	FALSE	NO
Sunny	Cool	Normal	FALSE	YES
Rainy	Mild	Normal	FALSE	YES
Sunny	Mild	Normal	TRUE	YES
Overcast	Mild	High	TRUE	YES
Overcast	Hot	Normal	FALSE	YES

B) A multi-class confusion matrix with three class labels A, B, and C is given in Table 2. Evaluate the classification accuracy, precision, recall and F1 metrics for each class. (3)

Table 2. Confusion matrix



		PREDICTED			
		A	B	C	SUM
ACTUAL	A	100	30	40	170
	B	20	120	60	200
	C	60	30	90	180
	SUM	180	180	190	550

C) How the data preparation methods like Binning and Normalization help in this process? Explain. (2)

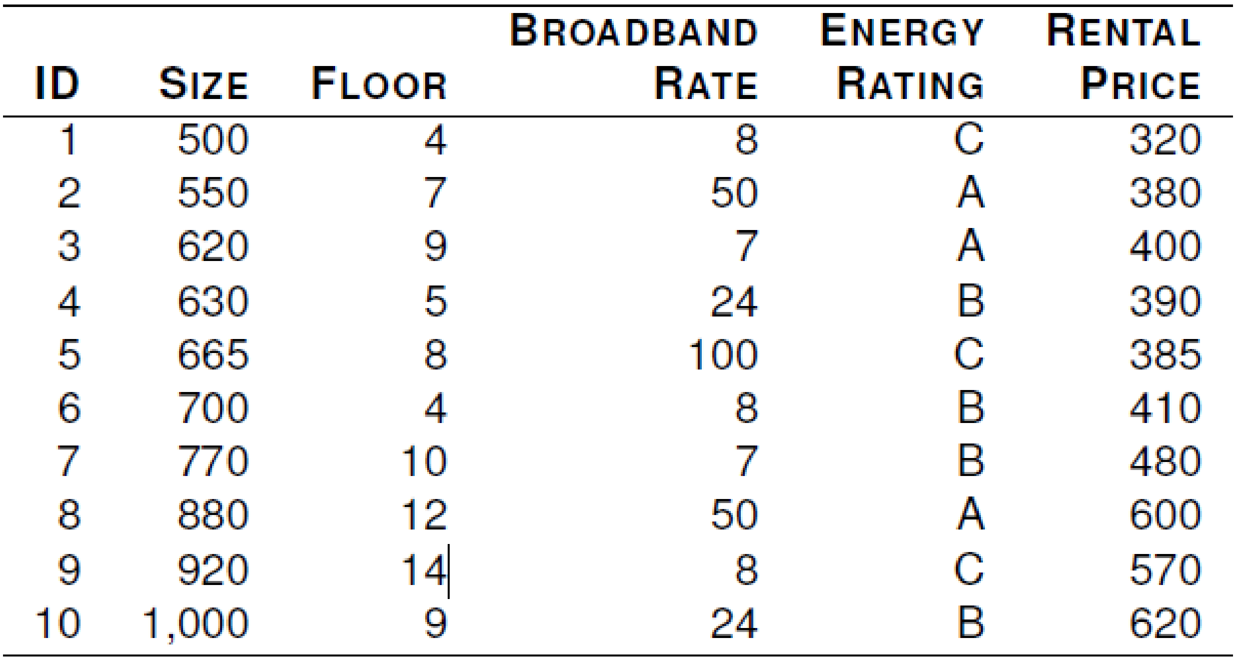
2) A) A machine is built to make mass-produced items. Each item made by the machine has a probability p of being defective. Given the value of p, the items are independent of each other. Because of the way in which the machines are made, p could take one of several values. In fact, p = X/100 where X has a discrete uniform distribution on the interval [0, 5]. The machine is tested by counting the number of items made before a defective is produced. Find the conditional probability distribution of X given that the first defective item is the thirteenth to be made. (5)

B) What is the multi-variate linear regression method? How hyperplanes are helpful in the learning process? (3)

C) How are the concepts of neurons and perceptrons realized in machine learning model developments? (2)

3) A) The office rental data set in a particular city XYZ is shown in Table 3. The size of the rental space in meter square, floor number, energy rating and the associated broadband rate are indicated. Rental Price is assumed to in the current currency of the city XYZ. (5)

Table 3. Office rental data in a city XYZ.



			Broadband	Energy	Rental
ID	Size	Floor	Rate	Rating	Price
1	500	4	8	C	320
2	550	7	50	A	380
3	620	9	7	A	400
4	630	5	24	B	390
5	665	8	100	C	385
6	700	4	8	B	410
7	770	10	7	B	480
8	880	12	50	A	600
9	920	14	8	C	570
10	1,000	9	24	B	620

Build a simple linear regression model to compute the best rental price by only considering the feature "Size". Use initial weights w[0] = 6.47, w[1] = 0.62 and learning rate = 0.00002. Compute the first three iterations and list the updated weights. Calculate the Error after each iteration.

B) How decision trees are used in the predictive analysis. Explain the concept of the ID3 algorithm. (3)

C) Explain the concept of pruning and boosting used while model learning. Discuss whether the early stopping concept in tree pruning is a good strategy to overcome overfitting. (2)

4) A) Cluster the following eight points (with (x, y) representing locations) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9) with the initial cluster centers, C1, C2 and C3 are assigned with A1(2, 10), A4(5, 8) and A7(1, 2) respectively. Using the distance metric identify the cluster centers after four iterations. (5)

B) How cross-validation enhances the evaluation during the model performance evaluation. Explain the concept of k-Fold Cross Validation. (3)

C) Discuss the k-nearest neighbour algorithm method as a similarity-based prediction model. (2)

5) A) How does Data Visualization help in the initiation of data model building? With examples explain the following data visualization methods that can be adopted (5)

a. Histogram Bar Plots

b. Scatter Plots

c. Box Plots

d. Contour Plots

e. Strips and Jitter Plots.

B) Describe the Logistic Linear regression. Use the data from Table 1 and demonstrate the logistic regression algorithm. (3)

C) How does the basic structure of the multivariable linear regression model allow the modelling of categorical? Demonstrating the algorithm using the data from Table 3. (2)

-----End-----