MACHINE LEARNING TOOLS AND TECHNOLOGIES-ICT 4304

OPEN ELECTIVE -II

Date : 1-01-2022

Q1. The Table Q1 contains data about stolen vehicles. Build a Naïve Bayes Classifier model and Classify the sample (Red, Sports, Imported)

Table Q1

Sl.No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No

(5)

Q2. Consider the data from a questionnaires survey and objective testing with two attributes (acid durability and Strength) to classify whether a special paper tissue is good or not. Four training samples are listed in table Q2. The factory has recently produced a new paper tissue that pass laboratory test with X1= 3 and X2= 7. Without another expensive survey ,Classify this new tissue using KNN.

TABLE . Q2

X1= Acid Durability(s)	X2 = Strength(Kg/m^2)	Y= Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

(3)

Q3. Classify the following activities under supervised or unsupervised learning. State the appropriate reasons for the same.

i. Identifying whether a lump is malignant or benign based on standard data sample taken from repository

ii. Establishing an appropriate algorithm that would identify research groups working in various domain (2)

Q4. A company wants to be able to discriminate between Buyers and Non Buyers based on the following characteristics: Age ∈{>40,31-40, <30}, Income ∈{Medium, Low, High}, Employed ∈{Yes, No}, Credit ∈{Fair, Excellent}. The training data is given in table Q4. Learn a decision tree using the ID3 /CART algorithm and draw the tree . Consider the decision tree to predict whether a person will buy a computer or not.

Table Q4

Sl. No.	Age	Income	Employed	Credit	Buy
1	<30	High	No	Fair	NO
2	<30	High	No	Excellent	No
3	31-40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31-40	Low	yes	Excellent	Yes
8	<30	Medium	No	Fair	No
9	<30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	<30	Medium	Yes	Excellent	Yes
12	31-40	Medium	No	Excellent	Yes
13	31-40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

(5)

Q5. We have a test dataset (table Q5) of 10 records with expected outcomes and a set of predictions from our classification algorithm. Evaluate the model in terms of accuracy, F1 score & Specificity.

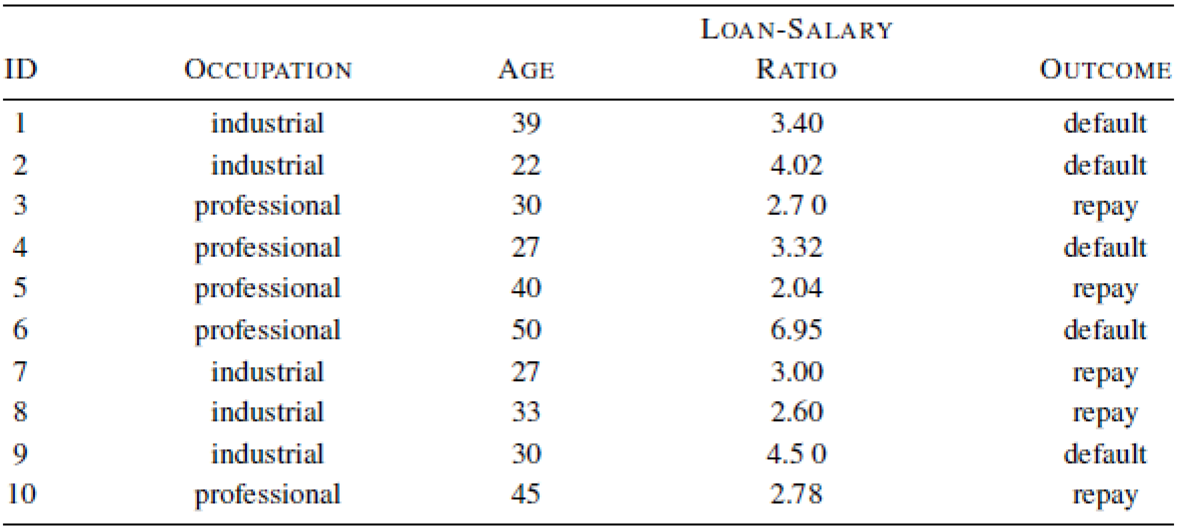
Table Q5

Sl. No.	**Expected**	**Predicted**
1	Man	Woman
2	Man,	Man
3	Woman	Woman
4	Man	Man
5	Woman	Man
6	Woman	Woman
7	Woman	Woman
8	Man	Man
9	Man	Woman
10	Woman	Woman

3)

Q6. The table Q6 lists a dataset from the credit scoring domain. We list two prediction models (Model 1 & Model 2) that are consistent with this dataset.

Table Q 6



ID	Occupation	Age	Loan-Salary Ratio	Outcome
1	industrial	39	3.40	default
2	industrial	22	4.02	default
3	professional	30	2.7 0	repay
4	professional	27	3.32	default
5	professional	40	2.04	repay
6	professional	50	6.95	default
7	industrial	27	3.00	repay
8	industrial	33	2.60	repay
9	industrial	30	4.5 0	default
10	professional	45	2.78	repay

Model 2:	**Model 1:**

If Age = 50 then
Outcome = Default
Else if Age = 39 then
Outcome = Default
Else if Age =30 and Designation = Senior then
Outcome = Default
Else if Age =27 and Designation = Junior then
Outcome = Default
Else
Outcome = Repay
	If Loan-Salary Ration > 3.00 then
	Outcome = Default
	Else
	Outcome = Repay

i. Which of the given models(Mode1/ Model 2) do you think will generalize better to the data instances not contained in the dataset?

ii. Whether the model you have rejected in Question **6 (i)** is overfitted or underfitted ? Explain. (2)