VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

DIGITAL ASSIGNMENT -2
ITA5007 – DATA MINING AND BUSINESS INTELLIGENCE
WINTER SEMESTER 2021-22
B2 SLOT

Note: Answer the questions in A4 sheet, scan and upload it as digital assignment.

Construct a decision tree using ID3 algorithm for the following sunburn dataset. The classification problem is binary and the dependent variable could be identified in the last column of the data set. The first column, name of the person which has no bearing on the outcome. Therefore it can be ignored while processing the data. The remaining attributes, Hair, Height, Weight and Lotion, which are nominal in nature could be used to construct decision tree.

Name	**Hair**	**Height**	**Weight**	**Lotion**	**Result**
Sarah	Blonde	Average	Light	No	Sunburned
Dana	Blonde	Tall	Average	Yes	None
Alex	Brown	Short	Average	Yes	None
Annie	Blonde	Short	Average	No	Sunburned
Emily	Red	Average	Heavy	No	Sunburned
Pate	Brown	Tall	Heavy	No	None
John	Brown	Average	Heavy	No	None
Katie	Blonde	Short	Light	Yes	None

The following table shows the training dataset pertains to bank load applications. The target class label is ‘Risk Class’. Using Naïve Bayes classifier, predict the class label for the test sample, {Yes, No, Female, Yes, A}

Owns Home	**Married**	**Gender**	**Employed**	**Credit Rating**	**Risk Class**
Yes	Yes	Male	Yes	A	B
No	No	Female	Yes	A	A
Yes	Yes	Female	Yes	B	C
Yes	No	Male	No	B	B
No	Yes	Female	Yes	B	C
No	No	Female	Yes	B	A
No	No	Male	No	B	B
Yes	No	Female	Yes	A	A
No	Yes	Female	Yes	A	C
Yes	Yes	Female	Yes	A	C

In the following IRIS dataset, independent attributes are Sepal length and Sepal width. The target class label is Species. Using KNN Classifier, classify the test sample with the attribute values {Sepal length = 5.2, Sepal width = 3.1} using 5 nearest neighbours.

Sepal length	**Sepal width**	**Species**
5.3	3.7	Setosa
5.1	3.8	Setosa
7.2	3.0	Virginica
5.1	3.4	Setosa
5.4	3.3	Setosa
5.1	3.9	Setosa
7.4	2.8	Virginica
6.1	2.8	Verscicolor
7.3	2.9	Virginica
6.0	2.7	Verscicolor
5.8	2.8	Virginica
6.3	2.3	Verscicolor
5.1	2.5	Verscicolor
6.3	2.5	Verscicolor
5.5	2.4	Verscicolor

The following dataset contains monthly e-commerce sales and the online advertising costs for 7 online stores during last year. Using Least squares method based linear regression analysis, find the equation of the straight line that fits the data best.

Online Store	**Monthly E-Commerce Sales (in 1000s)**	**Online advertising cost (in 1000s)**
1	368	1.7
2	340	1.5
3	665	2.8
4	954	5
5	331	1.3
6	556	2.2
7	376	1.3

The classification performance of a binary classifier with 80/20 splitting is summarized in the below contingency table.

Actual Class /Predicted Class	Disease = Yes	Disease = No
Disease = Yes	90	210
Disease = No	140	9560

Calculate TPR,TNR,FPR,FNR and Accuracy