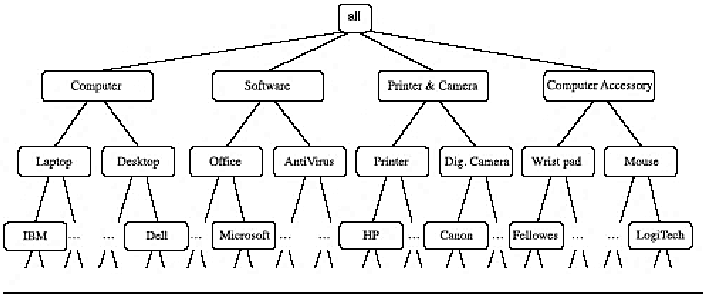
WALCHAND COLLEGE OF ENGINEERING

(Government Aided Autonomous Institute)
Visharambag, Sangli – 416415

Final Year B.Tech. Information Technology
ESE , ODD SEMESTER, AY 2023-24
Data Mining (5IT401)

ESE

MODEL ANSWERS

Q1	**A)**	State the methods to fill in the missing values for attributes in data mining process. Ans- •Ignore the tuple: usually done when class label is missing. •Fill in the missing value manually: tedious + infeasible? •Use a global constant to fill in the missing value: e.g., “unknown”, a new •Use the attribute mean to fill in the missing value •Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter •Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree	3	CO1
	B)	What is a box plot? Mention the two conditions that represent the outliers. Find first and third quartile for following data 25,28,29,29,30,34,35,35,37,38 Ans- A box plot is a special type of diagram that shows the quartiles in a box and the line extending from the lowest to the highest value. Outliers are greater than Q3+(1.5. IQR) or less than Q1-(1.5. IQR) The first quartile is the middle value of the lower half of the data, and it is represented by Q1. The third quartile is the middle value of the upper half of the data and is represented by Q3. first quartile 29, Third quartile: 35	3	CO2
	C)	Explain various approaches for mining multilevel association rules with reduced minimum support at lower levels. Elaborate one approach for following example.  all Computer Software Printer & Camera Computer Accessory Laptop Desktop Office AntiVirus Printer Dig. Camera Wrist pad Mouse IBM Dell Microsoft HP Canon Fellowes LogiTech Ans- • uniform minimum support threshold for all levels • reduced minimum support at lower levels 1. level by level independent 2. level cross filtering by k-itemset 3. level cross filtering by single-item 4. controlled level cross filtering by single-item ▪ The given concept hierarchy has 5 levels, referred to as levels 0 to level 4. ▪ Starting with level 0 at the root node for all. ▪ Here, level 1 includes computer, software, printer and camera so on…. ▪ Level 2 includes laptop computer, desktop computer and so on…. ▪ Level 3 includes IBM desktop computers…. ▪ Level 4 is the most specific abstraction level of this hierarchy which includes raw data.	4	CO3

Q2 **A)** State the advantages and disadvantages of ‘Decision Tree classification’.

Ans-

Advantages.

- Able to generate understandable rules
- Handles numerical and categorical data
- Clear indication of imp fields for prediction

Disadvantages

- Some trees deal with binary values
- Eror prone if no of training examples are less per class
- Growing tree process is expensive

2

CO1

B) What are the difficulties will arise when a decision tree is constructed?

Ans-

Guillotine cut- decision tree examines 1 dimension at time. If attribute is numeric

Overfit -decision tree T is overfitted if there is other tree T’ which is simplification of T,such that T has smaller error tha T’ over training set but T’ has smaller error than T over entire data

Attribute selection error- Wrong attribute selection for splitting at higher levels.

2

C) A training data is given as follows. Find overall entropy and info gain for attribute - Outlook and humidity.

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	NoPlay
sunny	hot	high	true	NoPlay
overcast	hot	high	false	Play
rain	mild	high	false	Play
rain	cool	normal	false	Play
rain	cool	normal	true	NoPlay
overcast	cool	normal	true	Play
sunny	mild	high	false	NoPlay
sunny	cool	normal	false	Play
rain	mild	normal	false	Play
sunny	mild	normal	true	Play
overcast	mild	high	true	Play
overcast	hot	normal	false	Play
rain	mild	high	true	NoPlay

ANS-

Info(T)= entropy (9/14,5/14) (play/total,noplay/total)=0.94

Info(outlook,T)=5/14 info(3/5,2/5)+ 4/14 info(1,0)+ 5/14 info(3/5,2/5)

=5/14(-3/5log3/5-2/5log2/5+5/14(-3/5log3/5-2/5log2/5=0.694

Gain(outlook,T)=info(T)-info(outlook,T)=0.94-0.694=0.246

Info(humidity,T)=

Gain(humidity,T)= 0.151

4

CO3

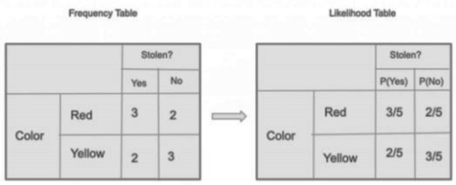
D) Use Naïve Bayes Algorithm to predict class for following case

X= {Color=Red; Type=SUV; Origin=Domestic; Stolen=?}

Example	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Blue	Sports	Domestic	No
5	Blue	Sports	Imported	Yes
6	Blue	SUV	Imported	No
7	Blue	SUV	Imported	Yes
8	Blue	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

5

CO3



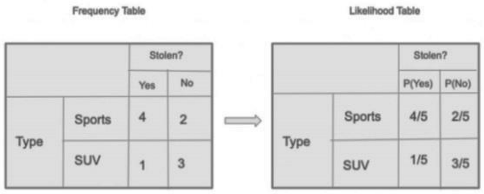
Frequency Table

		Stolen?	
		Yes	No
Color	Red	3	2
	Yellow	2	3

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Color	Red	3/5	2/5
	Yellow	2/5	3/5

Frequency and Likelihood tables of 'Color'



Frequency Table

		Stolen?	
		Yes	No
Type	Sports	4	2
	SUV	1	3

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Type	Sports	4/5	2/5
	SUV	1/5	3/5

Frequency and Likelihood tables of 'Type'

P(Yes|x) = 3/5*1/5*2/5/1/5=0.024
P(No|x)=2/5*3/5*3/5*1/2=0.07

Since 0.07 > 0.02, Which means given the features RED SUV and Domestic, our example gets classified as 'NO' the car is not stolen.

Q3 **A)** State the typical requirements of clustering in data mining. — 3 — CO1

Ans-

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

B) Write a short note on - Grid-based clustering method. — 3 — CO2

The grid-based clustering methods use a multi-resolution grid data structure. It quantizes the object areas into a finite number of cells that form a grid structure on which all of the operations for clustering are implemented. The benefit of the method is its quick processing time, which is generally independent of the number of data objects, still dependent on only the multiple cells in each dimension in the quantized space.

STING is a grid-based multiresolution clustering method in which the spatial area is divided into rectangular cells. There are generally several levels of such rectangular cells corresponding to multiple levels of resolution, and these cells form a hierarchical mechanism each cell at a high level is separation to form several cells at the next lower level. Statistical data regarding the attributes in each grid cell (including the mean, maximum, and minimum values) is precomputed and stored.

Statistical parameters of higher-level cells can simply be calculated from the parameters of the lower-level cells. These parameters contain the following: the attribute-independent parameter, count, and the attribute-dependent parameters, mean, stdev (standard deviation), min (minimum), max (maximum); and the type of distribution that the attribute value in the cell follows, including normal, uniform, exponential, or none (if the distribution is anonymous).

The statistical parameters can be used in top-down, grid-based approaches as follows. First, a layer within the hierarchical architecture is decided from which the query-answering procedure is to start. This layer generally includes a small number of cells. For every cell in the current layer, it can compute the confidence interval (or estimated range of probability) reflecting the cell's relevancy to the given query.

Figure….

C) With reference to DBSCAN algorithm define following terms and also draw diagram to represent them- minPts, eps, Core point, Border point — 3 — CO2

- minPts: The minimum number of points (a threshold) clustered together for a region to be considered dense.

- eps (ε): A distance measure that will be used to locate the points in the neighborhood of any point.
- Core point — This is a point that has at least *m* points within distance *n* from itself.
- Border point — This is a point that has at least one Core point at a distance *n*.

D) What is Dendrogram? Apply Hierarchical clustering Technique with single linkage to form clusters of following data. Draw resulting Dendrogram. CO3

City/ Distance	A	B	C	D	E	F
A	0	662	877	255	412	996
B	662	0	295	468	268	400
C	877	295	0	754	564	138
D	255	468	754	0	219	869
E	412	268	564	219	0	669
F	996	400	138	869	669	0

Ans-

A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or clusters are broken up (top-down view).

5

	A	B	C/F	D	E
A	0	662	877	255	412
B	662	0	295	468	268
C/F	877	295	0	754	564
D	255	468	754	0	219
E	412	268	564	219	0

Step 1

	A	B	C/F	D/E
A	0	662	877	255
B	662	0	295	268
C/F	877	295	0	564
D/E	255	268	564	0

Step 2

	A/D/E	B	C/F
A/D/E	0	268	564
B	268	0	295
C/F	564	295	0

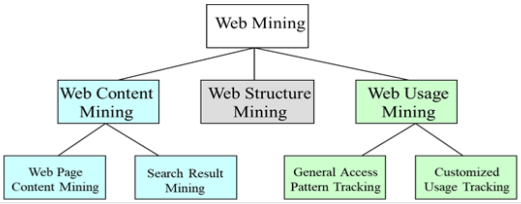
Step 3

	A/B/D/E	C/F
A/B/D/E	0	295
C/F	295	0

Step 4

Q4 **A)** Write a short note on – Web mining and its major types with figure CO2

Ans



3

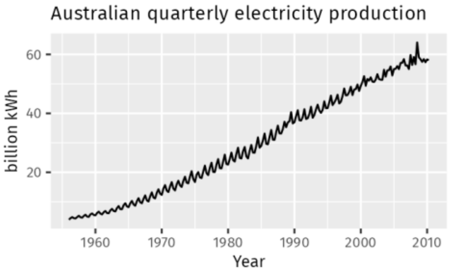
B) Discuss in detail- Spatial Classification and Spatial Trend Analysis CO3

Ans

A. Spatial classification

i. Analyze spatial objects to derive classification schemes, such as decision trees in relevance to certain spatial properties (district, highway, river, etc.)
Example: Classify regions in a province into *rich* vs. *poor* according to the

3

		average family income B. Spatial trend analysis i. Detect changes and trends along a spatial dimension ii. Study the trend of nonspatial or spatial data changing with space Example: Observe the trend of changes of the climate or vegetation with the increasing distance from an ocean		
	C)	Discuss in detail – mining from image data and give a few applications of image mining Image data: i. Extracted by aggregation and/or approximation ii. Similarity search in image data 1. Image sample based queries 2. Image feature specification queries iii. Size, color, shape, texture, orientation, and relative positions and structures of the contained objects or regions in the image Applications Medical science imaging Security, CBIR Etc	3	CO3
	D)	Define – Trend, Seasonal and Cyclic in time series data. Give your comment on following figure for existence of 'Trend, Seasonal and Cyclic' behaviour.  Trend A *trend* exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend as "changing direction", when it might go from an increasing trend to a decreasing trend. Seasonal A *seasonal* pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and known frequency. Cyclic A *cycle* occurs when the data exhibit rises and falls that are not of a fixed frequency. Figure shows a strong increasing trend, with strong seasonality. There is no evidence of any cyclic behaviour	4	CO2
		· · · · · *End of question paper* · · · · ·		