USN											CMRIT

Internal Assessment Test 3 – July 2022
Scheme of Evaluation

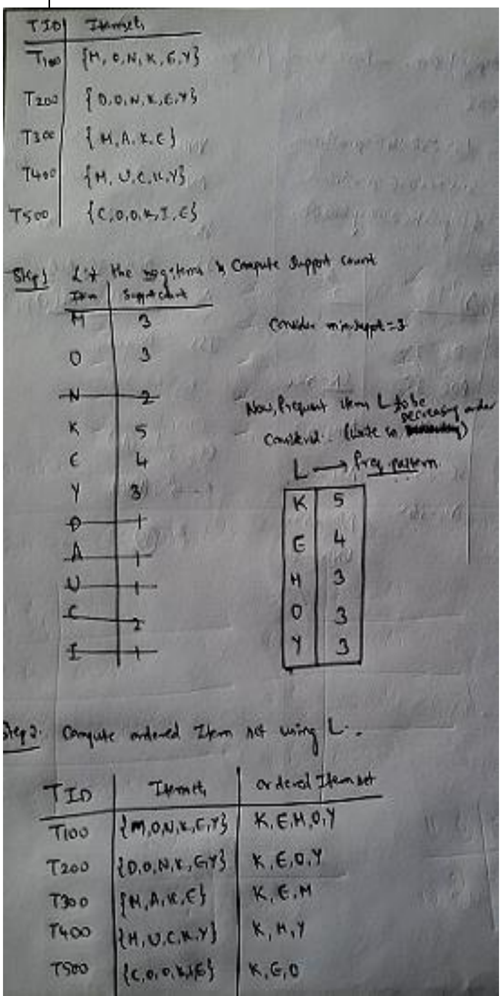
Sub:	DATA MINING AND DATA WAREHOUSING				Sub Code:	18CS641	Branch:	ISE
Date:	11/07 /2022	Duration:	90 min's	Max Marks: 50	Sem/Sec:	VI / A, B & C		OBE

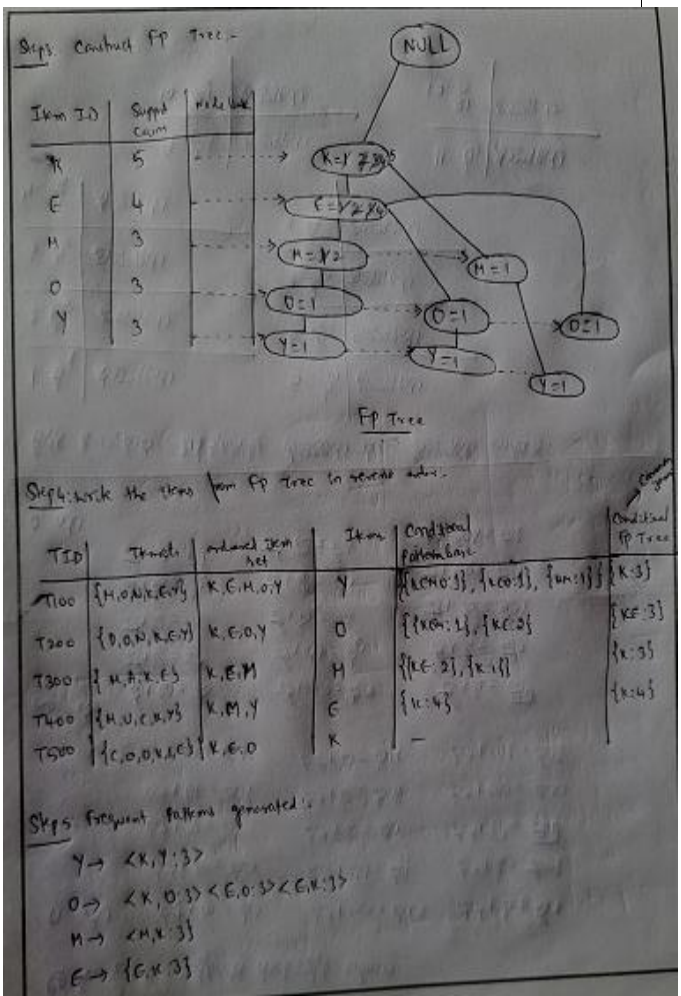
	Answer any FIVE FULL Questions	MARKS	CO	RBT
1 (a)	Construct the FP tree and generate the frequent item set using FP growth algorithm.	05	CO3	L3

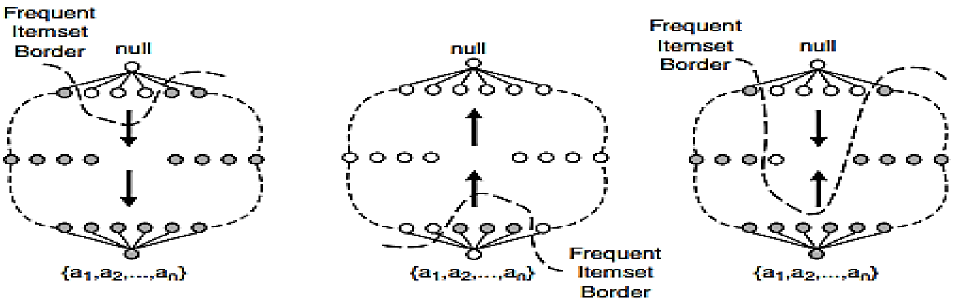
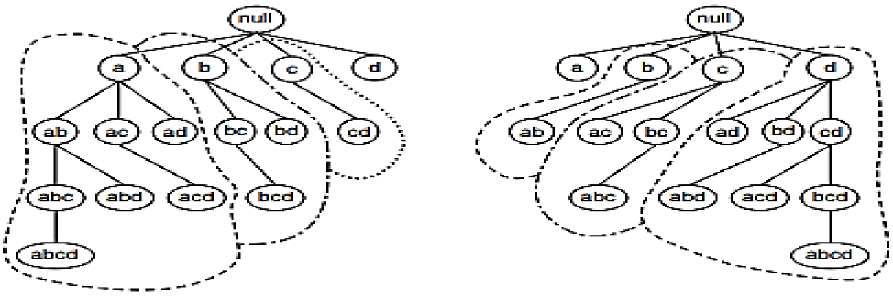
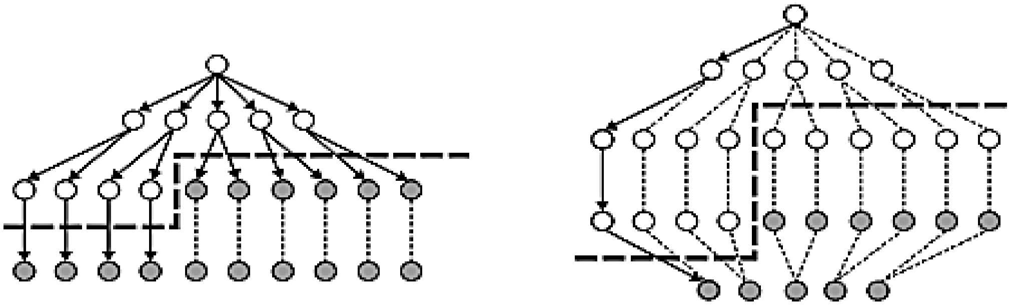
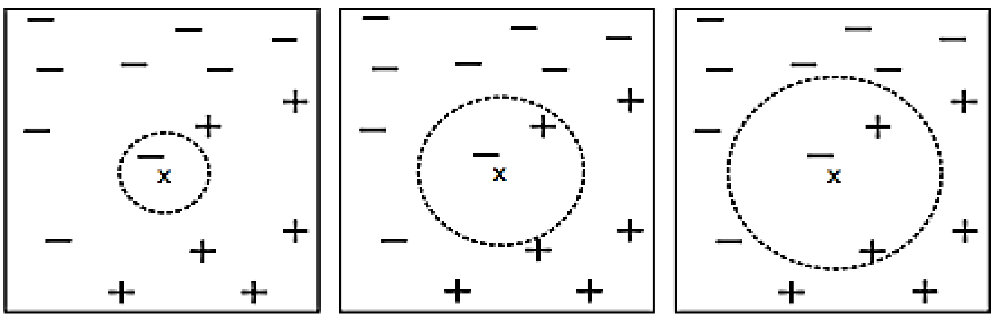
TID	items bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

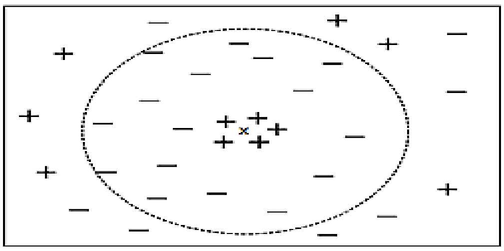
Scheme: Computing FP Tree and Frequent Item Set Carries **3+2** Marks.

Solution:





1 (b)	Explain alternate methods for Generating frequent itemsets. **Scheme:** Explanation of General-to-Specific versus Specific-to-General, Equivalence Classes, Breadth-First versus Depth-First:**2**+**1**+**2** Marks. **Solution:**  (a) General-to-specific (b) Specific-to-general (c) Bidirectional  (a) Prefix tree. (b) Suffix tree.  (a) Breadth first (b) Depth first	05	CO3	L2
2	Explain the concept and characteristics of K Nearest Neighborhood classifier with example. **Scheme:** Explanation with diagrams and examples, Algorithm and Characteristics carries **5+5** Marks **Solution:**  (a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor	10	CO4	L2



k-nearest neighbor classification with large k.

Algorithm 5.2 The k-nearest neighbor classification algorithm.

```
Let k be the number of nearest neighbors and D be the set of training examples.
for each test example z = (x', y') do
   Compute d(x', x), the distance between z and every example, (x, y) ∈ D.
   Select D_z ⊆ D, the set of k closest training examples to z.
   y' = argmax_v Σ_{(x_i, y_i) ∈ D_z} I(v = y_i)
end for
```

The characteristics of the nearest-neighbor classifier are:

• Nearest-neighbor classification is part of a more general technique known as instance-based learning, which uses specific training instances to make predictions without having to maintain an abstraction (or model) derived from data.

• Lazy learners such as nearest-neighbor classifiers do not require model building.

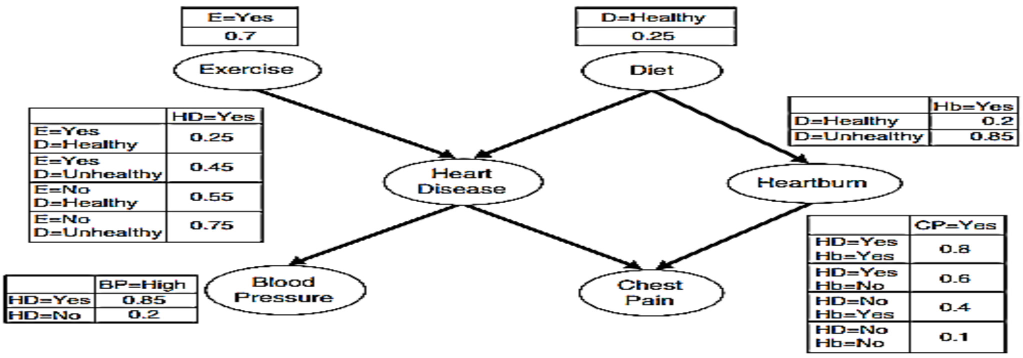
• Nearest-neighbor classifiers make their predictions based on local information, whereas decision tree and rule-based classifiers attempt to find a global model that fits the entire input space.

• Nearest-neighbor classifiers can produce arbitrarily shaped decision boundaries.

• Nearest-neighbor classifiers can produce wrong predictions unless the appropriate proximity measure and data preprocessing steps are taken.

3 Apply Bayesian Belief network for the given network and perform the following :

a. No Prior Information b. High BP c. High BP, Healthy Diet & Regular Exercise



Scheme: Computing for all the 3 cases carries **3+3+4** Marks 10 CO4 L3

Solution:

$$\begin{aligned} P(\text{HD} = \text{Yes}) &= \sum_{\alpha}\sum_{\beta} P(\text{HD} = \text{Yes}|E = \alpha, D = \beta)P(E = \alpha, D = \beta) \\ &= \sum_{\alpha}\sum_{\beta} P(\text{HD} = \text{Yes}|E = \alpha, D = \beta)P(E = \alpha)P(D = \beta) \\ &= 0.25 \times 0.7 \times 0.25 + 0.45 \times 0.7 \times 0.75 + 0.55 \times 0.3 \times 0.25 \\ &\quad + 0.75 \times 0.3 \times 0.75 \\ &= 0.49. \end{aligned}$$

Since P (HD = no) =1 – P (HD = yes) =0.51, the person has a slightly higher chance of not getting the disease.

	$P(\mathtt{BP}=\mathtt{High}) = \sum_{\gamma} P(\mathtt{BP}=\mathtt{High}\mid\mathtt{HD}=\gamma)P(\mathtt{HD}=\gamma)$ $= 0.85\times 0.49 + 0.2\times 0.51 = 0.5185.$ $P(\mathtt{HD}=\mathtt{Yes}\mid\mathtt{BP}=\mathtt{High}) = \frac{P(\mathtt{BP}=\mathtt{High}\mid\mathtt{HD}=\mathtt{Yes})P(\mathtt{HD}=\mathtt{Yes})}{P(\mathtt{BP}=\mathtt{High})}$ $= \frac{0.85\times 0.49}{0.5185} = 0.8033.$ Similarly, P(HD = No|BP = High)=1 − 0.8033 = 0.1967. Therefore, when a person has high blood pressure, it increases the risk of heart disease. $P(\mathtt{HD}=\mathtt{Yes}\mid\mathtt{BP}=\mathtt{High}, D=\mathtt{Healthy}, E=\mathtt{Yes})$ $= \left[\frac{P(\mathtt{BP}=\mathtt{High}\mid\mathtt{HD}=\mathtt{Yes}, D=\mathtt{Healthy}, E=\mathtt{Yes})}{P(\mathtt{BP}=\mathtt{High}\mid D=\mathtt{Healthy}, E=\mathtt{Yes})}\right]$ $\times P(\mathtt{HD}=\mathtt{Yes}\mid D=\mathtt{Healthy}, E=\mathtt{Yes})$ $= \frac{P(\mathtt{BP}=\mathtt{High}\mid\mathtt{HD}=\mathtt{Yes})P(\mathtt{HD}=\mathtt{Yes}\mid D=\mathtt{Healthy}, E=\mathtt{Yes})}{\sum_{\gamma} P(\mathtt{BP}=\mathtt{High}\mid\mathtt{HD}=\gamma)P(\mathtt{HD}=\gamma\mid D=\mathtt{Healthy}, E=\mathtt{Yes})}$ $= \frac{0.85\times 0.25}{0.85\times 0.25 + 0.2\times 0.75}$ $= 0.5862,$ The probability that the person does not have heart disease is P(HD = No|BP = High, D = Healthy, E = Yes)=1 − 0.5862 = 0.4138. The model therefore suggests that eating healthily and exercising regularly may reduce a person’s risk of getting heart disease.			
4	Explain rule ordering schemes and how a rule-based classifier works. **Scheme:** Explanation of How a Rule-Based classifier works and Rule-ordering schemes with examples carries **5+5** Marks **Solution:** ➢ Mutually Exclusive Rules ➢ Exhaustive Rules ➢ Ordered Rules ➢ Unordered Rules **Rule-Based Ordering** (Skin Cover=feathers, Aerial Creature=yes) ==> Birds (Body temperature=warm-blooded, Gives Birth=yes) ==> Mammals (Body temperature=warm-blooded, Gives Birth=no) ==> Birds (Aquatic Creature=semi)) ==> Amphibians (Skin Cover=scales, Aquatic Creature=no) ==> Reptiles (Skin Cover=scales, Aquatic Creature=yes) ==> Fishes (Skin Cover=none) ==> Amphibians **Class-Based Ordering** (Skin Cover=feathers, Aerial Creature=yes) ==> Birds (Body temperature=warm-blooded, Gives Birth=no) ==> Birds (Body temperature=warm-blooded, Gives Birth=yes) ==> Mammals (Aquatic Creature=semi)) ==> Amphibians (Skin Cover=none) ==> Amphibians (Skin Cover=scales, Aquatic Creature=no) ==> Reptiles (Skin Cover=scales, Aquatic Creature=yes) ==> Fishes	10	CO4	L2

5	Illustrate the concept of estimating a confidence interval for accuracy and Comparing the performance of two classifiers. **Scheme:** Explanation of concept of estimating a confidence interval for accuracy and Comparing the performance of two classifiers carries **5+5** Marks **Solution:** 1. The experiment consists of N independent trials, where each trial has two possible outcomes: success or failure. 2. The probability of success, p, in each trial is constant. $$P(X=v)=\binom{N}{p}p^v(1-p)^{N-v}.$$ For example, if the coin is fair ($p=0.5$) and is flipped fifty times, then the probability that the head shows up 20 times is $$P(X=20)=\binom{50}{20}0.5^{20}(1-0.5)^{30}=0.0419.$$ If the experiment is repeated many times, then the average number of heads expected to show up is $50\times0.5=25$, while its variance is $50\times0.5\times0.5=12.5$. $$P\Big(-Z_{\alpha/2}\leq\frac{acc-p}{\sqrt{p(1-p)/N}}\leq Z_{1-\alpha/2}\Big)=1-\alpha,$$ $$\frac{2\times N\times acc+Z^2_{\alpha/2}\pm Z_{\alpha/2}\sqrt{Z^2_{\alpha/2}+4Nacc-4Nacc^2}}{2(N+Z^2_{\alpha/2})}.$$ Let Mij denote the model induced by classification technique Li during the jth iteration. Note that each pair of models M1j and M2j are tested on the same partition j. Let e1j and e2j be their respective error rates. The difference between their error rates during the jth fold can be written as dj = e1j − e2j. $$\widehat{\sigma}^2_{d^{cv}}=\frac{\sum_{j=1}^k(d_j-\overline{d})^2}{k(k-1)},\qquad d^{cv}_t=\overline{d}\pm t_{(1-\alpha),k-1}\widehat{\sigma}_{d^{cv}}.$$	10	CO4	L2
6	Define Clustering. Explain K-means as an optimization Problem using SSE and SAE. **Scheme:** Defining Clustering and K-means using SSE and SAE Carries **5+5** Marks. **Solution:** An entire collection of clusters is commonly referred to as a clustering. **Derivation of K-means as an Algorithm to Minimize the SSE** In this section, we show how the centroid for the K-means algorithm can be mathematically derived when the proximity function is Euclidean distance and the objective is to minimize the SSE. Specifically, we investigate how we can best update a cluster centroid so that the cluster SSE is minimized. In mathematical terms, we seek to minimize Equation 8.1, which we repeat here, specialized for one-dimensional data. $$\text{SSE}=\sum_{i=1}^{K}\sum_{x\in C_i}(c_i-x)^2 \qquad (8.4)$$	10	CO5	L2

Here, C_i is the i^{th} cluster, x is a point in C_i, and c_i is the mean of the i^{th} cluster. See Table 8.1 for a complete list of notation.

We can solve for the k^{th} centroid c_k, which minimizes Equation 8.4, by differentiating the SSE, setting it equal to 0, and solving, as indicated below.

$$\begin{aligned} \frac{\partial}{\partial c_k}\text{SSE} &= \frac{\partial}{\partial c_k}\sum_{i=1}^{K}\sum_{x\in C_i}(c_i - x)^2 \\ &= \sum_{i=1}^{K}\sum_{x\in C_i}\frac{\partial}{\partial c_k}(c_i - x)^2 \\ &= \sum_{x\in C_k} 2*(c_k - x_k) = 0 \end{aligned}$$

$$\sum_{x\in C_k} 2*(c_k - x_k) = 0 \Rightarrow m_k c_k = \sum_{x\in C_k} x_k \Rightarrow c_k = \frac{1}{m_k}\sum_{x\in C_k} x_k$$

Thus, as previously indicated, the best centroid for minimizing the SSE of a cluster is the mean of the points in the cluster.

Derivation of K-means for SAE

To demonstrate that the K-means algorithm can be applied to a variety of different objective functions, we consider how to partition the data into K clusters such that the sum of the Manhattan (L_1) distances of points from the center of their clusters is minimized. We are seeking to minimize the sum of the L_1 absolute errors (SAE) as given by the following equation, where $dist_{L_1}$ is the L_1 distance. Again, for notational simplicity, we use one-dimensional data, i.e., $dist_{L_1} = |c_i - x|$.

$$\text{SAE} = \sum_{i=1}^{K}\sum_{x\in C_i} dist_{L_1}(c_i, x) \qquad (8.5)$$

We can solve for the k^{th} centroid c_k, which minimizes Equation 8.5, by differentiating the SAE, setting it equal to 0, and solving.

$$\begin{aligned} \frac{\partial}{\partial c_k}\text{SAE} &= \frac{\partial}{\partial c_k}\sum_{i=1}^{K}\sum_{x\in C_i}|c_i - x| \\ &= \sum_{i=1}^{K}\sum_{x\in C_i}\frac{\partial}{\partial c_k}|c_i - x| \\ &= \sum_{x\in C_k}\frac{\partial}{\partial c_k}|c_k - x| = 0 \end{aligned}$$

$$\sum_{x\in C_k}\frac{\partial}{\partial c_k}|c_k - x| = 0 \Rightarrow \sum_{x\in C_k} sign(x - c_k) = 0$$

If we solve for c_k, we find that $c_k = median\{x \in C_k\}$, the median of the points in the cluster. The median of a group of points is straightforward to compute and less susceptible to distortion by outliers.

Faculty Signature CCI Signature HOD Signature