1)

a) i) A core object in DBSCAN algorithm with respect to eps and MinPts is referred to as the object which has more objects than MinPts within the eps radius, or the ε neighbourhood.

A border object in DBSCAN algorithm with respect to eps and MinPts is referred to as the object which has less objects than MinPts within the radius eps or the ε neighbourhood, but is a neighbour of a core object, i.e. is density-reachable from atleast one core object.

An object that is neither a core object nor a border object is a noise object. A noise object is not density-reachable from any other core object.

4½

ii) The points p and q are said to be density-reachable to q if there exists a sequence of points p1, p2, ..., pn, where p1 = p, and pn = q, such that pi+1 is directly density reachable from pi, for all of the points i.

Two points p and q are said to be density connected to each other wrt ε and MinPts if there exists a point o such that both p and q are density-reachable from point o through a sequence.

b)

i) Maximality condition of the DBSCAN algorithm states that if $p \in C$, (a cluster) & from a set of points in D and q is another point in D which is density reachable from point C, then point $q \in C$, i.e. the same cluster.

ii) Connectivity condition of the DBSCAN algorithm states that if p and q are two points in D, such that $p, q \in C$, then the points p and q are density connected to each other in the same cluster.

2

c) There are two parameters Eps and MinPts of the DBSCAN algorithm which can be determined by finding the value of K, or the no of clusters. For each object and its K^{th} neighbour, is also known as the K-distance. The K-distance between an object and its K^{th} neighbour, is also known as the K-distance. Now, for the core objects, the value of K-distance will be small if the cluster is small and the K^{th} neighbour is inside the cluster. Also, for the border objects, the value of the K-distance will be very large. For each object, its K-distance is found out & and then the sorted values of the K-distance are plotted in a graph. A point in the graph where there is a sharp change in the value, or an inflection determines the eps value for a particular K.

3½

MinPts

The value of k gives us the value of the [redacted] for the algorithm. Several values of k are tried and a particular value of k is finally selected, such that it gives the proper number of clusters. If the value of k selected is very high, then a large number of noise objects will be treated as clusters. Also, if the value of k selected is very small, then a small compact cluster will be treated as noise. So, the value of k is chosen very carefully, so that the value of MinPts is successfully found out.

2) a) i) Purity refers to the maximum weight of an object (r_i) in the cluster i divided by the size of the cluster C.

Mathematically, purity $= \frac{1}{n_i} \cdot \max(n_{ij})$, where $n_j \in C$

The value of purity is biased, because it attains the maximum value when the total number of clusters is n.

For example, if there is a cluster which contains 5, 2 and 1 objects respectively, then the purity of the cluster would be $= \frac{1}{8} \max\{(5, 2, 1)\} = 5/8$.

$= 0.625$

iii) Rand Index is an external validation criterion which assess how many samples are correctly clustered by the clustering algorithm, with the help of ground truth.

Mathematically, RI (Rand Index) =

$$\frac{\text{No of samples clustered correctly}}{\text{Total no of samples present}}$$

For example, if for an example the following table is created,

3

	No of samples as A in ground truth truth	No of samples as B in the ground truth
No of samples as A in the cluster	70	60
No of samples as B in the cluster	50	30

Then the Rand Index for the following sample will be-

$$RI = \frac{70+30}{70+60+50+30} = \frac{100}{210}$$

$$= 0.476$$

b) Intercluster distance between two objects X_i and X_j is defined if the two objects belong to two different clusters. It is represented by $\delta(X_i, X_j)$

Intracluster distance between two objects X_i and X_j is defined if the two objects belong to the same cluster. It is represented by $\Delta(X_K)$.

The two intercluster distances are as follows-

i) Single linkage distance - It refers to the minimum distance between two objects belonging to two different clusters.

Mathematically $\delta_1(X_i, X_j) = \min \{ dist(X_i, X_j) \}$ $X_i \in S, X_j \in T$

4½

ii) Complete linkage distance - It refers to the distance between the two most remote objects belonging to two different clusters. Mathematically, $\delta_2(X_i, X_j) = \max \{ dist(X_i, X_j) \}$ $X_i \in S, X_j \in T$.

The two intracluster distances are as follows-

i) Complete diameter distance- It refers to the maximum distance between two objects belonging to the same cluster.

Mathematically, $\Delta_1(X_i, X_j) = \max \{ dist(X_i, X_j) \}$ $X_i, X_j \in S$.

ii) Average diameter distance - It refers to the average of all the distances between two objects that are present in the same cluster.

Mathematically, $\Delta_2(X_i, X_j)$

$$\Delta_2(X_K) = \frac{1}{|S|\,|S-1|} \sum_{\substack{X_i, X_j \in S \\ i \neq j}} dist(X_i, X_j)$$

c) The Dunn's Index of D-Index for cluster validation is defined as

$$DIndex(U) = \min_{i=1}^{n} \left\{ \min_{\substack{j=1, \\ j \neq i}}^{n} \left\{ \left\{ \frac{\delta(X_i, X_j)}{\max\{\Delta(X_K)\}} \right\} \right\} \right\}$$

(3)

The Dunn's Index is an important index as it compares both the intercluster and intracluster distances. Here $\delta(X_i, X_j)$ is the intercluster distance and $\Delta(X_K)$ denotes the intracluster distance. The main aim of Dunn's index is to maximise the intercluster distance and minimise the intracluster distance, so that the compactness and connectedness is more between two within the cluster and the spatial separation is more between the two clusters. The more the value of the Dunn's Index, the more the cluster quality. Thus, Dunn's Index helps in the usefulness of cluster evaluation.

3) a) The support of an association rule of the form $X \rightarrow Y$, is denoted by $(X \cup Y)$. count is the number of times $(X \cup Y)$ appears in the transaction divided by the total no of transactions. If the no of transactions is n. then support $= \frac{(X \cup Y).\text{count}}{n}$

The confidence of an association rule of the form $X \rightarrow Y$, is the no of transactions where if X occurs, then Y will occur.

Mathematically, confidence $(X \rightarrow Y) = \frac{(X \cup Y).\text{count}}{X.\text{count}}$

4) An itemset is referred to as a frequent itemset if the support of the itemset is greater than the minimum support, i.e. support $(X) \geq \text{minSup}$.

An association rule is referred to as an important rule if follows these three conditions.

i) confidence $(X \rightarrow Y) \geq \text{min.Conf}$

ii) Support $(X \rightarrow Y) =$ support $(X \cup Y) =$ support (Z)

iii) confidence $(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}$

b) The transactions defined are as follows-

Transactions	Items
T1	Bread, Butter
T2	Bread, Milk, Butter
T3	Bread, Jelly, Butter
T4	Bread, Coke
T5	Bread, Milk
T6	Milk, Coke

According to apriori algorithm we have to find the frequent item-sets.

Now let C_1 = {Bread}, {Butter}, {Milk}, {Jelly}, {Coke}

Now, support (Bread) = 5/6 = 0.83

support (Butter) = 3/6 = 0.5

support (Milk) = 3/6 = 0.5

support (Jelly) = 1/6 = 0.17

support (Coke) = 2/6 = 0.33

Now, minSup = 30% = 0.3.

So, F_1 : {Bread}, {Butter}, {Milk}, {Coke}

Now, let C_2 : {Bread, Butter}, {Bread, Milk}, {Bread, Coke}, {Butter, Milk}, {Butter, Coke}, {Milk, Coke}

Now, support (Bread Butter) = 3/6 = 0.5
support (Bread, Milk) = 2/6 = 0.33
support (Bread, Coke) = 1/6 = 0.17
support (Butter, Milk) = 1/6 = 0.17
support (Butter, Coke) = 0
support (Milk, Coke) = 1/6 = 0.17

So, F_2: {Bread, Butter}, {Bread, Milk}

Now, C_3: {~~Bread, Butter, Milk~~} ∅
(No other candidates are generated from F_2).

So, the frequent itemset = {Bread, Butter}, and {Bread, Milk}.

5

Now, for the generation of important association rules, we may have the following -

i) Bread → Butter, confidence = 3/5 = 60%
ii) Butter → Bread, confidence = 3/3 = 100%
iii) Bread → Milk, confidence = 2/5 = 40%
iv) Milk → Bread, confidence = 2/3 = 66.6%

Now, minimum confidence = 80%
So only association rule satisfies this condition.
So, the important association rule is -
Butter → Bread.

c) If there are m items available then the total number of transaction rules could be of the order, $O(2^m)$. Since this grows exponentially the cost of computational and memory requirements are very high. The apriori algorithm exploits the sparseness of the data and high support and confidence. Still the apriori algorithm generates thousands, and even millions of association rules. This is a major drawback of apriori algorithm as the computational cost and cost of memory becomes very high.

2

6) a) The decision tree model trained by the dataset is as follows-

4)

a) The dataset is given as follows.

Outlook	Humidity	Wind	Play Tennis (Class Variable)
Sunny	Normal	Strong	Yes
Sunny	High	Weak	No
Overcast	High	Strong	Yes
Rain	Normal	Strong	No
Rain	High	Weak	Yes
Sunny	Normal	Weak	Yes
Sunny	High	Strong	No
Rain	High	Strong	No
Rain	Normal	Weak	Yes
Overcast	Normal	Weak	Yes

Total no of samples, $N = 10$

Now, the information required would -

$$\text{Info}(D) = \frac{-p}{p+n}\log\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log\left(\frac{p}{p+n}\right)$$

$$= -\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right)$$

$$= -\frac{3}{2}\log_2\left(\frac{3}{2}\right) - 1\log_2(1)$$

$$= -\frac{3}{2}\left(\log_2(3) - \log_2(2)\right) - 1\times 0$$

$$= -\frac{3}{2}\left($$

$$= -\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right)$$

$$= -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$

$$= -\frac{3}{5}\left(\log_2(3) - \log_2(5)\right) - \frac{2}{5}\left(\log_2(2) - \log_2(5)\right)$$

$$= -\frac{3}{5}(1.6 - 2.3) - \frac{2}{5}(1 - 2.3)$$

$$= \frac{3}{5} \times 0.7 + \frac{2}{5} \times 1.3 = 0.42 + 0.52 = 0.94.$$

Now, for each of the attributes, we consider the information and information gain.

1) Outlook -

$$\text{Info}_{\text{Outlook}}(D) = \sum_{i=1}^{v} \frac{|D_v|}{|D|} \text{Info}(D_v)$$

$$\text{Info}(\text{Sunny}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= \frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= \frac{1}{2}\log 2 + \frac{1}{2}\log_2 2 = \log_2 2 = 1.$$

$$\text{Info}(\text{Rain}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1.$$

$$\text{Info}(\text{Overcast}) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - 0$$

$$= 0.$$

$$\text{So, Info}_{\text{Outlook}}(D) = \frac{4}{10} \times 1 + \frac{4}{10} \times 1 + \frac{2}{10} \times 0$$

$$= \frac{8}{10} = 0.8.$$

$\therefore$ Gain (Outlook) = 0.14

ii) Humidity -

Info (Normal) $= -\frac{4}{5}\log_2(4/5) - \frac{1}{5}\log_2(1/5)$

$= -\frac{4}{5}\left(\log_2(4) - \log_2(5)\right) - \frac{1}{5}\left(\log_2(1) - \log_2(5)\right)$

$= -\frac{4}{5} \times \left(2 - 2.3\right) - \frac{1}{5}\left(0 - 2.3\right)$

$= \frac{4}{5} \times 0.3 + \frac{1}{5} \times 2.3 = 0.24 + 0.46 = 0.7$

Info (High) $= -\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5) = 0.94$

$\therefore$ Info Humidity (D) $= \frac{5}{10} \times 0.7 + \frac{5}{10} \times 0.94$

$= 0.35 + 0.47 = 0.82$

$\therefore$ Gain (Humidity) $= 0.94 - 0.82 = 0.12$

4+2

iii) Wind -

Info (Strong) $= -\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5) = 0.94$

Info (Weak) $= -\frac{4}{5}\log_2(4/5) - \frac{1}{5}\log_2(1/5)$

$= 0.7.$

$\therefore$ Info Wind (D) $= \frac{5}{10} \times 0.94 + \frac{5}{10} \times 0.7 = 0.35 + 0.47$

$= 0.82$

$\therefore$ Gain (Wind) $= 0.94 - 0.82 = 0.12$

So, among the 3 attributes, gain of attribute, Outlook is the highest, so the root of the decision tree will the attribute Outlook.

b) The possible terminating criteria of a decision tree algorithm are -

i) When all of the nodes of the decision tree have the same class label. The decision tree stops growing and the leaf node is assigned the value of the class node.

ii) When none of the attributes are left for splitting a node into further nodes. In such cases, a majority decision is taken for assigning the class label to the leaf node.

iii) When all of the data samples are exhausted. In such cases each leaf node is assigned a class label according to the value of the node.

(3)

c) Over-fitting happens when a model tries to fit all of the data including noise and outliers into itself, resulting in very low error rates in the training data set but high error rates in the unseen data.

Over-fitting also occurs when all of the training data is used for model construction resulting in very low accuracy for the test data.

(3)

In decisions trees, the problem of overfitting can be solved in two ways-

i) Pre-pruning- In pre-pruning, the growth of the tree is stopped, if a split to the node does not increase the goodness measure beyond a certain threshold. This threshold is generally determined by the user.

ii) Post pruning- In post-pruning, the decision tree is let to fully grow and then those branches are removed which reflects the noise and outliers data. In such cases, the decision tree removes the more complex decisions which may be influenced by noisy data.