DATE:

Assignment 2

1) Explain different metrics used for evaluating classifier performance.

Ans i) Accuracy -
This is the most basic measure for classifier performance, and simply measures the proportion of correct predictions made by the classifier.

ii) Precision -
This metric measures the proposition of true positives out of all instances predicted to be positive by the classifier.

iii) F1 score -
This is the harmonic mean of the precision and recall, and provides a way to balance these two metrics.

iv) Confusion matrix -
This is a matrix that shows the number of true positives, true negatives false positives and false negatives for a classifier and provides a more detailed view of its performance.

2) What is linear regression? Explain how it is different from classification.

Ans Linear regression is one of the easiest and most popular machine learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression make predictions for

continous/real or numeric variables such as sales, age, product prices etc.

	Regression	Classification.
1)	In Regression, the objective output variable must be of continous nature or real value.	In classification, the output variable must be discrete value.
2)	Used for continous data	Used for discrete data.
eg	Linear regression	Logistic regression.

3) Differentiate agglomerative and divisive clustering algorithm.

Parameters	Agglomerative	Divisive.
Category	Bottom-up approach	Top-down approach.
Approach	Each data in its own cluster and the algorithm recursively merges the closest pairs of clusters until a single cluster containing all points is obtained.	All data points are in a single cluster, and the algorithm recursively splits the cluster into smaller subsets until each point is in its own cluster.
Complexity	It is more complex.	It is less complex comparatively

DATE:

Outliers	Can handle them better than divisive clustering.	Divisive may create sub-clusters around outliers.

4) Explain two feature selection measures in building decision tree.

ans i) Information gain (IG) -

It is a widely used feature selection measure in decision tree algorithms. It is based on the concept of entropy, which measures impurity or randomness of a set of data. IG is calculated as the reduction in entropy achieved by splitting the data on a particular feature. The higher the IG the more informative the feature is considered to be.

IG = Entropy before split - (Entropy after split)

Gini index -

It is another commonly used feature selection method. It is similar to entropy. It is calculated as the probability of misclassification of a randomly chosen element in the dataset when using the feature for splitting.

$$Gini = 1 - \Sigma(p)^2$$

p is probability of each class.

5) What are the different measures for finding intercluster distance? Determine the distance between two clusters a(17, 42, 10) and c2(20, 36, 8) using a single linkage minimum distance techniques.

DATE:

ans. i) Euclidean distance -
This is the most commonly used distance metric, and it calculates the straight distance between two points.

ii) Manhattan distance -
Also known as the L1 distance or city block distance. It calculates the sum of the absolute differences between the co-ordinates of two points in a grid-like fashion.

iii) Minkowski distance -
Minkowski distance is a generalised distance metric that includes Euclidean distance and Manhattan distance as special cases.

$c_1 (17, 42, 10)$ $c_2 (20, 36, 38)$

$$\text{Distance} = \sqrt{(20-17)^2 + (36-42)^2 + (10-8)^2}$$
$$= \sqrt{3^2 + (-6)^2 + 2^2}$$
$$= \sqrt{9+4+36}$$
$$= 7$$

6) Explain Agglomerative (AGNES) clustering algorithm with an example. Comment on Dendogram and cluster formation process.

ans It is a hierarchical clustering algorithm that iteratively merge similar clusters based on a chosen similarity or dissimilarity metric until a stopping condition is met. The algorithm starts with all points in ~~one cluste~~

DATE:

its own cluster and then successively merges clusters until a single cluster containing all points is formed.

eg (A, B, C, D, E, F).

i) Initialization:
All points forming its own cluster in start
(A), (B), (C), (D), (E), (F).

ii) Compute similarity / dissimilarity-
Calculate similarity / dissimilarity between clusters based on a chosen metric.

iii) Merge clusters-
Merge two most similar clusters into a single cluster. Let's say (A) & (B) have the highest similarity, so they merge into a new cluster (AB).

iv) Update the similarity-
Recalculate the similarity between newly formed clusters (AB), (C), (D), (E), and (F)

v) Repeat step 2-4 till only one cluster is left in the end.

vi) Dendrogram-
During clustering a tree like structure is formed. It shows the hierarchical structure of the clusters and the order in which they

were merged

vii) Cluster formation -
The final clusters are formed by cutting the dendogram at a certain height or distance, which corresponds to desired number of clusters.

Q7) A database has five transactions. Let min sup = 50% and min confidence = 50%. Find all frequent algorithms using Apriori algorithm.

TID	Item_bought
T1	{M, O, N, K, E, Y}
T2	{D, O, N, K, E, Y}.
T3	{A, K, E}.
T4	{C, O, O, K, I, E}.
T5	{M, U, C, K, E}.

support = 50% = $\frac{50}{100} \times 5 = 3$.

1-itemset -

Item	frequency.
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2

DATE:

1- item set = { E, K, M, O, Y }

2 - item set -

item	frequency
E K	4
E M	2
E O	3
E Y	2
K M	3
K O	3
K Y	3
M O	1
M Y	2
O Y	2

2 - item set = { ~~{E, K}~~ {E, K}, {E, O}, {K, M}, {K, O}, {O, Y} }

3 - item set -

item	frequency
E, K, O	3
E, K, M	2
E, K, Y	2
E, O, M	1
E, O, Y	2
K, M, O	1
K M, Y	2
K, O, Y	2

3 - item set = { { E, K, O } }

DATE:

8) A dataset has 5 transactions. Let min sup and min confidence = ~~50%~~ 80 %. Find all frequent itemsets using FP growth.

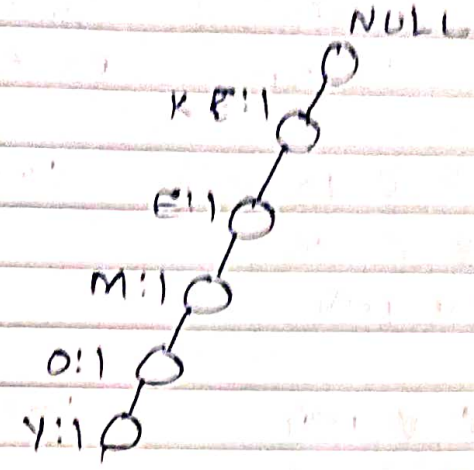
TID	Item_bought.
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, E}

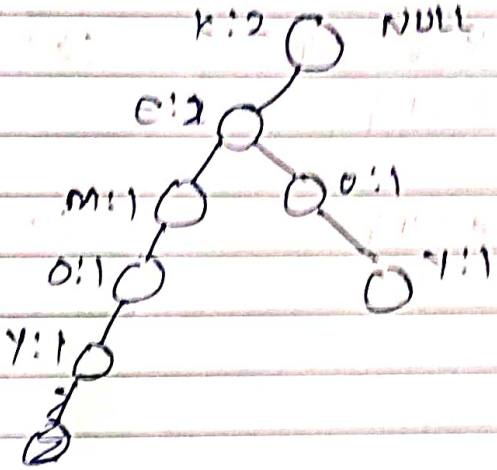
Items	coont.
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	4
U	1
Y	3

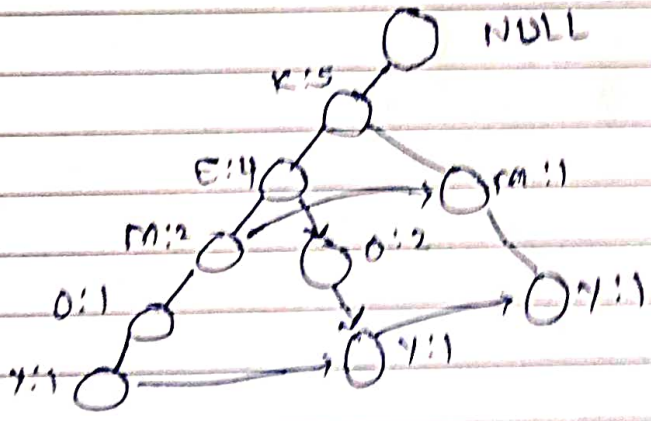
1-item set = {E, K, M, O, Y}.

TID	List of items.
T100	{E, K, M, O, Y}
T200	{E, K, O, Y}
T300	{E, K, M}
T400	{K, M, Y}
T500	{E, K, O}

DATE:







Making FP tree -

DATE:

Itemset	Conditional Pattern	Conditional FP-tree	Frequent Pattern
{Y}.	{K, E, M, O:1} {K, E, O:1}. {K, M:1}.	{K:3}	{K, Y:3}.
{O}	{K, E, M:1} {K, E:2}	{K:3, E:3}	{K, O:3} {E, O:3}
{M}	{K, E:2} {K:1}	{K:3}	{K, M:3}
{E}	{K:4}	{K:4}	{K, E:4}
{K}	-	-	-