DETAILED DATA MINING ANSWERS – POINTWISE FORMAT

1. KDD Process (Knowledge Discovery in Databases)

1.1 Data Selection – Identify relevant data from databases, warehouses. Example: Choosing customer transactions.

1.2 Data Cleaning – Remove noise, errors, missing values. Example: Replace missing ages with averages.

1.3 Data Integration – Combine multiple data sources. Example: Merge online & offline sales.

1.4 Data Transformation – Normalize, aggregate, encode data. Example: Convert categories to numbers.

1.5 Data Mining – Apply algorithms like Apriori, clustering, classification.

1.6 Pattern Evaluation – Filter meaningful patterns using support, confidence, etc.

1.7 Knowledge Presentation – Use graphs, dashboards to present results.

2. Data Mining Architecture

2.1 Data Sources – Databases, warehouses, log files.

2.2 Data Warehouse – Stores cleaned & historical data.

2.3 Pre-processing Module – Cleans, transforms, reduces data.

2.4 Data Mining Engine – Performs association, classification, clustering, regression.

2.5 Pattern Evaluation – Calculates support, confidence, gini index, etc.

2.6 GUI – Allows user interaction and visualization.

3. FP-Growth Algorithm

3.1 Overview – Mines frequent itemsets without candidate generation.

3.2 Steps – Scan database → Build FP-tree → Mine conditional FP-trees.

3.3 Advantages – Faster than Apriori, low memory usage.

3.4 Limitations – Large FP-tree for sparse datasets.

4. Improving Efficiency of Apriori Algorithm

4.1 Hash-Based Technique – Reduces candidate itemsets using hash tables.

4.2 Transaction Reduction – Removes irrelevant transactions.

4.3 Partitioning – Divides database into partitions.

4.4 Sampling – Uses small subsets for faster mining.

4.5 Dynamic Itemset Counting – Adds candidates during scanning.

5. Bayesian Belief Network

5.1 Definition – Probabilistic graphical model (DAG).

5.2 Components – Nodes, edges, CPT tables.

5.3 Working – Applies Bayes’ theorem to update probabilities.

5.4 Advantages – Handles uncertainty, complex relations.

5.5 Applications – Medical diagnosis, fraud detection, ML.

6. Backpropagation Algorithm

6.1 Definition – Trains neural networks using gradient descent.

6.2 Steps – Forward pass → Error calculation → Backward pass → Weight update.

6.3 Features – Uses chain rule, reduces error gradually.

7. KNN Algorithm

7.1 Definition – Instance-based method for classification/regression.

7.2 Steps – Choose K → Compute distance → Select neighbors → Vote/average.

7.3 Advantages – Simple, no training needed.

7.4 Limitations – Slow on big data, sensitive to outliers.

8. Vectors & Lists in R

8.1 Vectors – Same-type elements. Example: c(1,2,3).

8.2 Lists – Mixed data types. Example: list(name="Afthab").

9. Matrices, Arrays, Data Frames in R

9.1 Matrices – 2D same-type structures.

9.2 Arrays – Multi-dimensional structures.

9.3 Data Frames – Table with mixed-type columns.

10. Graphs in R

Line plot: plot(x,y,type="l")

Bar plot: barplot(values)

Histogram: hist(data)

Scatter: plot(x,y)

Pie chart: pie(values)