

VU LEARNING HUB

vulearninghub.xyz

HIGHLIGHTED HANDOUT

CS626

Data Mining Techniques

CS626 by VU Learning HUB

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>

CS626

Data Mining Techniques

This handout covers the complete course topic by topic in simple, exam-focused language. Each chapter begins with learning objectives and ends with key points and exam pointers. Definitions are boxed in green-friendly format, examples are worked out step by step, and very important exam points are flagged separately. Study the highlighted points carefully before your papers.



Course Information	
Course Code	CS626
Course Title	Data Mining Techniques
Credit Hours	3
Level	BS/MS

Table of Contents

Table of Contents	2
1. Introduction to Data Mining	3
2. Data Preprocessing	4
3. Data Warehousing and OLAP	6
4. Mining Frequent Patterns and Association Rules	8
5. Classification: Decision Trees and Bayesian Methods	10
6. Classification: Advanced Methods	12
7. Cluster Analysis: Partitioning Methods	14
8. Cluster Analysis: Hierarchical and Density Methods	16
9. Outlier and Anomaly Detection	18
10. Mining Text, Web and Sequential Data	19
11. Data Mining Applications and Trends	21
Glossary of Key Terms	23
Practice Questions for Exams	24
MCQ Practice Test	30
Exam Strategy Guide 	34

1. Introduction to Data Mining

Learning Objectives

- Define data mining and the KDD process
- Describe the kinds of data and patterns mined

Data mining is defined as the process of discovering interesting, non-trivial, previously unknown and potentially useful patterns from large amounts of data.

KDD (Knowledge Discovery in Databases) is defined as the overall process of extracting knowledge from data, of which data mining is the central step: selection -> preprocessing -> transformation -> data mining -> interpretation/evaluation.

- Why now: data is collected everywhere (transactions, web, sensors) but human analysis cannot keep up.
- Multidisciplinary: draws on databases, statistics, machine learning, visualization and high-performance computing.
- Data mining vs OLAP: OLAP summarizes and reports; data mining discovers hidden patterns and predicts.

Major data mining functionalities

- Characterization and discrimination: summarize a target class and contrast it with others.
- Frequent pattern mining: association and correlation (market basket analysis).
- Classification and prediction: build models from labelled data.
- Cluster analysis: group similar objects without labels.
- Outlier analysis: detect anomalies and fraud.

Key Points to Remember

- Data mining = discovery of hidden patterns; KDD = the full pipeline.
- Know the five functionalities by name.

Exam Pointer

Very important for exams: Define data mining and list the KDD steps — the opening question of most papers.

Data mining in daily life

- Recommendations on Daraz/YouTube, spam filters, fraud alerts on bank SMS.

- Invisible until it fails — then everyone notices.
- Every click trains someone's model.

CRISP-DM methodology

- Business understanding -> data understanding -> preparation -> modelling -> evaluation -> deployment.
- Iterative, not waterfall — expect to loop back.
- The industry-standard mining process model.

Data vs information vs knowledge

- Data: raw facts; information: processed data; knowledge: applied understanding.
- Mining climbs this ladder automatically.
- Wisdom — when to act — stays human.

Big data 5 Vs

- Volume, velocity, variety, veracity, value.
- Traditional tools choke; distributed systems cope.
- Start with value — big data without purpose is cost.

CRISP-DM phases

- Business understanding, data understanding, preparation, modelling, evaluation, deployment.
- The most cited methodology in data mining.
- It is iterative — loops back are normal.
- Agile variants suit modern teams.
- Examiners expect the six phases in order.

KDD vs data mining

- KDD is the full process; mining is one step.
- Fayyad's 1996 definition is the classic citation.
- Selection, preprocessing, mining, interpretation.
- The distinction appears in exam MCQs. 
- Use the terms precisely.

2. Data Preprocessing

Learning Objectives

- Explain why preprocessing is essential 
- Describe cleaning, integration, reduction and transformation

Real-world data is dirty: incomplete, noisy and inconsistent. Very important for exams: data preprocessing typically consumes 60-80% of the effort in a data mining project, and models built on dirty data produce misleading results ("garbage in, garbage out").

Data cleaning

- Missing values: ignore the tuple, fill manually, use the attribute mean/median/mode, or use the most probable value.
- Noisy data: binning (smooth by bin means/medians/boundaries), regression smoothing, clustering to detect outliers.

Data integration and transformation

- Integration: combine multiple sources; resolve entity identification, redundancy (correlation analysis) and value conflicts.
- Normalization: min-max scaling to [0,1], z-score normalization, decimal scaling — needed so large-valued attributes do not dominate distance measures.
- Aggregation, generalization and attribute construction create better features.

Min-max: $v_{\text{new}} = (v - \min) / (\max - \min) \times (\text{new_max} - \text{new_min}) + \text{new_min}$

Z-score: $v_{\text{new}} = (v - \text{mean}) / \text{std_dev}$

Data reduction

- Dimensionality reduction: remove irrelevant attributes (feature selection), PCA.
- Numerosity reduction: replace data by smaller representations (sampling, histograms, clustering).
- Discretization and concept hierarchies: convert continuous values to intervals.

Key Points to Remember

- Cleaning -> integration -> transformation -> reduction.
- Normalization is essential before distance-based mining.

Exam Pointer

Very important for exams: Numerical on min-max or z-score normalization is very common.

Data quality dimensions

- Accuracy, completeness, consistency, timeliness, believability.
- Poor quality costs more than cleaning does.
- Profile data before mining it.

Handling imbalanced data

- Fraud: 99.9% normal — accuracy lies; use F1/AUC.
- Fixes: resampling, SMOTE, class weights.
- Very important for exams: always check class balance first. 

ETL example

- Extract: pull sales from POS; Transform: clean, conform; Load: into warehouse nightly.
- Failures at 3am need alerting, not discovery at 9am.
- Incremental loads beat full refreshes.

Feature engineering

- Good features beat fancy algorithms.
- Domain knowledge creates the best features.
- Automate the boring, handcraft the crucial.

Data quality dimensions

- Accuracy, completeness, consistency, timeliness, validity.
- Poor quality costs more than cleaning does.
- Profile data before modelling.
- Business rules encode validity checks.
- Report quality metrics with every dataset.

Missing value strategies

- Delete rows only if missingness is tiny.
- Mean/median imputation is quick but biased.
- Model-based imputation preserves relationships.
- Missingness itself can be informative.
- Document the chosen strategy.

3. Data Warehousing and OLAP

Learning Objectives

- Define data warehouse and its architecture
- Explain OLAP operations

Data warehouse is defined as a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management decision-making.

- OLTP (operational databases): day-to-day transactions, detailed, current.
- OLAP (warehouses): analysis, summarized, historical.
- Architecture: data sources -> ETL (extract, transform, load) -> warehouse -> data marts -> OLAP tools.

Multidimensional data model

- Facts: numeric measures (sales amount). Dimensions: perspectives (time, product, location).
- Star schema: one central fact table, denormalized dimension tables.
- Snowflake schema: dimensions normalized into multiple tables. Fact constellation: multiple fact tables.

OLAP operations

- Roll-up: aggregate up a hierarchy (city -> country). Drill-down: reverse.
- Slice: select one dimension value; Dice: select a subcube.
- Pivot (rotate): change dimensional orientation.

Key Points to Remember

- Warehouse = integrated history for analysis; OLTP = live transactions.
- Star vs snowflake schema difference is a favourite question.

Exam Pointer

Very important for exams: Draw a star schema for a sales warehouse — a classic diagram question.

OLAP vs data mining example

- OLAP: "sales by region last quarter" — known question.
- Mining: "which customers will churn next month" — unknown answer.
- Use OLAP to explore, mining to predict.

Data marts vs warehouses

- Mart: departmental subset; warehouse: enterprise.
- Marts are faster to build and easier to own.
- Dependent marts keep definitions consistent.

Slowly changing dimensions

- Customers move, marry, rename — history must be kept.
- Type 2: new row per change; Type 1: overwrite.
- Type 2 preserves true history for analysis.

MOLAP vs ROLAP

- MOLAP: precomputed cubes, fast, less scalable.
- ROLAP: SQL on warehouse, flexible, slower.
- HOLAP mixes both.

OLAP operations

- Roll-up aggregates; drill-down disaggregates.
- Slice selects one dimension value; dice selects several.
- Pivot rotates the view.
- These five verbs answer most business questions.
- Practice each on a sample cube.

Star vs snowflake schema

- Star: denormalized, fast queries.
- Snowflake: normalized, less redundancy.
- Fact tables hold measures; dimensions hold context.
- Galaxy schemas share dimensions across facts.
- Draw both from memory for exams.

4. Mining Frequent Patterns and Association Rules

Learning Objectives

- Define support and confidence
- Explain the Apriori algorithm step by step

Association rule is defined as an implication of the form $A \rightarrow B$ meaning customers who buy A also tend to buy B, measured by support and confidence.

$$\text{support}(A \rightarrow B) = P(A \cup B) = \text{count}(A, B) / \text{total transactions}$$

$$\text{confidence}(A \rightarrow B) = P(B|A) = \text{count}(A, B) / \text{count}(A)$$

Example

Rule {bread} $\rightarrow$ {butter}: appears in 30 of 100 transactions, and bread appears in 50. Support = $30/100 = 30\%$. Confidence = $30/50 = 60\%$.

The Apriori algorithm

- Apriori property: all non-empty subsets of a frequent itemset must also be frequent — this prunes candidates.
- Step 1: find frequent 1-itemsets (count $\geq$ min support).
- Step 2: join frequent k-itemsets to form (k+1)-candidates; prune using the Apriori property.
- Step 3: scan the database to count candidates; repeat until no more frequent itemsets.
- Step 4: generate rules from frequent itemsets and keep those meeting min confidence.
- Limitation: many database scans and huge candidate sets — improved by FP-growth, which compresses data into an FP-tree and mines without candidate generation.

Key Points to Remember

- Support = frequency; confidence = strength; Apriori prunes with the subset property.
- FP-growth avoids candidate generation using the FP-tree.

Exam Pointer

Very important for exams: Worked Apriori numericals (find frequent itemsets, generate rules) are the most repeated questions.

Market basket in Pakistan retail

- Superstores use basket data for shelf placement and combos.
- Eid bundles are association rules made visible.
- Even small shops intuitively know "bread $\rightarrow$ eggs".

Lift and conviction

- Lift > 1 : positive association; < 1 : negative.
- Conviction measures rule directionality.
- Use lift to rank rules beyond confidence.

Sequential patterns example

- "Camera -> memory card -> tripod" across months.
- GSP mines such ordered patterns.
- Powers "complete your kit" recommendations.

Negative association rules

- "Buy X -> NOT buy Y" reveals cannibalization.
- Needs careful support thresholds.
- Useful in assortment planning.

Apriori pruning principle

- All subsets of a frequent itemset are frequent.
- Contrapositive: infrequent subsets kill supersets.
- This prunes the exponential search space.
- Still scans the database many times.
- FP-growth was invented to fix that.

Lift and conviction

- Lift above 1 means positive association.
- Conviction measures rule implication strength.
- Support filters noise; confidence filters weakness.
- All three appear in exam questions. ☰
- Compute them on a small example.

5. Classification: Decision Trees and Bayesian Methods

Learning Objectives

- Build a decision tree using information gain
- Explain Naive Bayesian classification

Classification is defined as supervised learning that builds a model from labelled training data to predict the class of new, unseen tuples.

Decision tree induction

- Each internal node tests an attribute, branches show outcomes, leaves give class labels.

- Attribute selection: choose the attribute with highest information gain (or gain ratio, or Gini index).
- Entropy measures impurity; information gain = reduction in entropy after splitting.
- Tree pruning (prepruning/postpruning) removes overfitting branches.

$$\text{Entropy}(S) = - \sum p_i \times \log_2(p_i) \text{ over classes } i$$

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum (|S_v|/|S|) \times \text{Entropy}(S_v)$$

Example

Training set: 9 yes / 5 no for "buys computer". Entropy = $-(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0.94$. Splitting on "age" gives the highest gain, so age becomes the root.

Bayesian classification

- Based on Bayes theorem: $P(H|X) = P(X|H) \times P(H) / P(X)$.
- Naive assumption: attributes are conditionally independent given the class — "naive" but fast and effective, especially for text.
- Handles missing values gracefully; needs only one scan of training data.

Key Points to Remember

- Decision trees: interpretable, need pruning. Naive Bayes: fast, probabilistic, independence assumption.
- Overfitting is the main enemy — prune trees and validate.

Exam Pointer

Very important for exams: Compute entropy and information gain for a small dataset — extremely common numerical.

Pruning example

- A tree that memorizes training data fails on new customers.
- Prune branches supported by few examples.
- Validate on holdout data — honesty over optimism.

C4.5 improvements over ID3

- Handles continuous attributes via split points.
- Gain ratio fixes multi-valued bias; pruning built in.
- Missing values handled gracefully.

Rule interestingness

- Support-confidence can mislead (milk -> bread confounds).
- Lift and chi-square filter spurious rules.
- Domain experts validate the survivors.

Cost-sensitive learning

- Missing fraud costs more than false alarms.
- Weight errors by business cost.
- Threshold tuning implements cost sensitivity.

Naive Bayes independence assumption

- Features are conditionally independent given the class.
- Rarely true, yet accuracy stays competitive.
- Needs little data and trains instantly.
- Laplace smoothing handles zero counts.
- A strong baseline for text classification.

Entropy and information gain

- Entropy measures impurity in bits.
- Gain = impurity before minus after split.
- Biased toward many-valued attributes.
- Gain ratio corrects that bias.
- Gini is the faster alternative.

6. Classification: Advanced Methods

Learning Objectives

- Explain kNN, neural networks and SVMs
- Describe model evaluation techniques

- k-Nearest Neighbours: lazy learner — stores all training data, classifies by majority vote of k nearest neighbours using a distance metric. Simple, no training, but slow at prediction.
- Neural networks: layers of neurons with weights learned by backpropagation; powerful for complex patterns but "black box" and slow to train.
- Support Vector Machines: find the maximum-margin hyperplane separating classes; excellent for high-dimensional data like text.

Evaluating classifiers

Metric	Formula	Meaning
Accuracy	$(TP+TN)/total$	Overall correctness
Precision	$TP/(TP+FP)$	Of predicted positives, how many are right
Recall	$TP/(TP+FN)$	Of actual positives, how many found
F1	$2PR/(P+R)$	Harmonic mean of precision and recall

- Confusion matrix organizes TP, TN, FP, FN.
- Holdout, cross-validation (k-fold) and bootstrap estimate true performance.
- ROC curves compare classifiers across thresholds.

Key Points to Remember

- Accuracy misleads on imbalanced data — use precision/recall/F1.
- Always evaluate on unseen data via cross-validation.

Exam Pointer

Very important for exams: Compute precision, recall and F1 from a confusion matrix — standard numerical.

Lazy vs eager learners

- Eager (trees, SVM): build model now, predict fast.
- Lazy (kNN): predict slow, no training.
- Choose by data size and prediction urgency.

Backpropagation intuition

- Error flows backward, adjusting guilty weights.
- Learning rate: too high diverges, too low crawls.
- Needs lots of data and patience.

Ensemble intuition

- Many weak models vote better than one strong one.
- Diversity among models is the secret ingredient.
- Random forests: the default first try.

Model deployment

- Export as API; monitor latency and drift.
- Version models like code.
- Rollback plans save careers.

Overfitting cures

- Pruning, cross-validation, regularization.
- More data beats cleverer models.
- Early stopping for neural networks.
- Ensembles average away variance.
- Always compare against a simple baseline.

ROC curves

- Plot TPR against FPR across thresholds.
- AUC summarizes ranking quality.
- Insensitive to class imbalance.
- Compare classifiers visually.
- Random guessing gives the diagonal.

7. Cluster Analysis: Partitioning Methods

Learning Objectives

- Explain k-means and k-medoids
- Discuss how to choose the number of clusters

Clustering is defined as unsupervised grouping of objects so that objects within a cluster are similar and objects across clusters are dissimilar.

k-Means

- 1. Choose k initial centroids. 2. Assign each point to the nearest centroid. 3. Recompute centroids as cluster means. 4. Repeat until stable.
- Strengths: simple, efficient, scalable. Weaknesses: must specify k, sensitive to outliers and initial centroids, works poorly on non-spherical clusters.
- k-medoids (PAM) uses actual data points as centres — more robust to outliers but costlier.

Choosing k and evaluating clusters

- Elbow method: plot within-cluster sum of squares vs k ; look for the "elbow".
- Silhouette coefficient measures how well each point fits its cluster (-1 to 1).
- Domain knowledge often guides the final choice.

Example

Customer segmentation: cluster shoppers by age and spending to find groups like "young high-spenders" for targeted marketing.

Key Points to Remember

- k-means = fast but needs k and hates outliers; k-medoids = robust but slower.
- Clustering has no labels — evaluation is harder than classification.

Exam Pointer

Very important for exams: Trace two iterations of k-means on a small 2-D dataset — a favourite numerical.

k-means in marketing

- Segment shoppers for targeted SMS campaigns.
- Telecoms cluster users for package design.
- Start with business-interpretable k .

Choosing k in k-means

- Elbow, silhouette, and business sense together.
- Try several k ; stability matters.
- Document why you chose it.

Clustering validation

- No labels: use internal indices and stability.
- External validation when partial labels exist.
- Business actionability is the real test.

Hierarchical clustering linkage

- Single: chaining; complete: compact; average: balanced.
- Choice shapes the dendrogram.
- Try two and compare.

Choosing k in k-means

- Elbow method: look for the bend.
- Silhouette scores measure cohesion.
- Domain knowledge beats heuristics.
- Try several k and inspect clusters.
- There is no universally correct k.

k-means limitations

- Assumes spherical, equal-size clusters.
- Sensitive to initialization and outliers.
- Needs k upfront.
- k-medoids and k-modes address variants.
- Know when NOT to use it.

8. Cluster Analysis: Hierarchical and Density Methods

Learning Objectives

- Explain agglomerative vs divisive clustering
- Describe DBSCAN and its parameters

- Agglomerative (bottom-up): each point starts as its own cluster; merge the closest pair repeatedly until one cluster remains.
- Divisive (top-down): start with one cluster and split.
- Linkage criteria: single (nearest points), complete (farthest), average, centroid.
- Result shown as a dendrogram; cut it at a chosen height to get clusters.

DBSCAN — density-based clustering

- Parameters: eps (neighbourhood radius) and minPts (minimum points to form a dense region).
- Core points have $\geq$ minPts neighbours; border points are reachable but not dense; others are noise/outliers.
- Finds arbitrarily shaped clusters and flags outliers automatically; struggles with varying densities.

Example

DBSCAN on GPS locations finds irregularly shaped "hotspot" regions of taxi pickups and marks isolated points as noise — k-means cannot do this.

Key Points to Remember

- Hierarchical gives a dendrogram; DBSCAN finds arbitrary shapes + outliers.
- DBSCAN needs eps and minPts tuning.

Exam Pointer

Very important for exams: Label core/border/noise points for a given eps and minPts — common short numerical.

Dendrogram reading

- Height shows merge distance — cut where jumps are large.
- Each leaf is a data point; branches are clusters.
- No need to pre-specify k.

DBSCAN in practice

- Tune eps via k-distance plot.
- Scales poorly to very high dimensions.
- Great for geo-spatial hotspots.

OPTICS extension

- Orders points by reachability — handles varying densities.
- Extract clusters at multiple levels.
- DBSCAN's sophisticated sibling.

Subspace clustering need

- High dimensions: clusters hide in subsets.
- CLIQUE finds dense subspaces.
- Relevant for gene expression data.

Dendrogram reading

- Height shows merge distance.
- Cut at a height to choose cluster count.
- Long vertical gaps suggest natural clusters.
- Leaf order is arbitrary — don't over-read it.
- Works best under a few hundred points.

DBSCAN parameters

- Eps: neighbourhood radius; MinPts: density threshold.
- k-distance plots guide Eps choice.
- Finds arbitrary shapes, ignores noise.
- Struggles with varying densities.
- OPTICS generalizes the idea.

9. Outlier and Anomaly Detection

Learning Objectives

- Define outliers and their causes
- Describe statistical, proximity and clustering-based detection

Outlier is defined as a data object that deviates significantly from the rest — generated by a different mechanism than the normal data.

- Causes: measurement errors, fraud, intrusions, rare events, or data entry mistakes.
- Applications: credit-card fraud, network intrusion, faulty sensors, medical anomalies.

Detection approaches

- Statistical: model normal data with a distribution; flag points with low probability (z-score > 3).
- Proximity-based: points far from their k nearest neighbours are outliers.
- Clustering-based: small clusters or points far from any cluster centre are suspicious.
- Density-based (LOF): compare local density of a point with its neighbours.

Key Points to Remember

- Outliers can be errors to remove or the most interesting discoveries (fraud).
- Always investigate before deleting.

Exam Pointer

Very important for exams: Give two applications and one detection method each — typical short question.

Fraud detection workflow

- Score transactions in real time; flag top 0.1%.

- Human analysts review flags — models suggest, people decide.
- Feedback retrains the model weekly.

Outlier treatment decisions

- Error -> correct or remove; interesting -> investigate.
- Never silently delete — document every removal.
- Robust methods reduce outlier influence.

Concept drift

- Fraud patterns evolve; models go stale.
- Monitor performance; retrain on schedules.
- Adaptive methods track drift automatically.

Anomaly vs novelty

- Anomaly: training had outliers; novelty: training was clean.
- One-class SVM suits novelty detection.
- Define which you face before choosing.

Distance-based outlier scores

- kNN distance flags isolated points.
- LOF compares local densities.
- Mahalanobis accounts for correlations.
- Normalize features first — scale distorts distance.
- Visualize before trusting scores.

Isolation Forest

- Outliers isolate in few random splits.
- Fast and scalable to big data.
- Few hyperparameters to tune.
- Scores normalize across the forest.
- A modern default for anomaly detection.

10. Mining Text, Web and Sequential Data

Learning Objectives

- Explain text mining steps
- Describe web mining types and sequential pattern mining

Text mining

- Pipeline: tokenization -> stop-word removal -> stemming/lemmatization -> vectorization (bag of words, TF-IDF).
- TF-IDF weights terms by frequency in the document vs rarity across documents.
- Applications: sentiment analysis, spam filtering, topic modelling.

Web mining

- Content mining: extract knowledge from page text.
- Structure mining: analyse hyperlinks (PageRank ranks pages by link importance).
- Usage mining: analyse clickstreams and logs to personalize sites.

Sequential patterns and time series

- Sequential pattern mining finds ordered patterns, e.g. customers who buy camera then memory card (GSP algorithm).
- Time-series: trend, seasonality and forecasting (moving averages).

Key Points to Remember

- TF-IDF = term frequency x inverse document frequency.
- Web mining splits into content, structure and usage.

Exam Pointer

Very important for exams: Compute TF-IDF for a term in a tiny corpus — increasingly asked.

Sentiment analysis for brands

- Track Twitter/Facebook sentiment during campaigns.
- Urdu-English code-mixed text is the local challenge.
- Aspect sentiment tells what exactly customers praise.

PageRank intuition

- A vote from an important page counts more.
- Damping factor models random jumps.

- Powers Google's original ranking.

Recommender systems

- Collaborative: "users like you liked"; content: "similar items".
- Cold start: new users/items need fallback strategies.
- Hybrid systems dominate production.

Text preprocessing pitfalls

- Aggressive stemming merges distinct meanings.
- Stop-word removal can hurt phrases.
- Validate preprocessing on samples.

PageRank intuition

- A page matters if important pages link to it.
- Random surfer model formalizes this.
- Damping factor handles dead ends.
- Power iteration computes it at scale.
- Google's founding algorithm.

Sequential pattern example

- Customers buying A then B then C.
- GSP and PrefixSpan mine such patterns.
- Time constraints refine the search.
- Web clickstreams are the classic data.
- Distinguish from time-series mining.

11. Data Mining Applications and Trends

Learning Objectives

- List real-world applications
- Discuss privacy, big data and future trends

- Retail: market basket analysis, recommendation systems, churn prediction.
- Banking: credit scoring, fraud detection, anti-money-laundering.
- Healthcare: disease prediction, drug discovery, outbreak detection.
- Telecom: churn analysis; education: learning analytics; government: tax fraud, security.

Challenges and trends

- Big data: volume, velocity, variety — needs distributed mining (MapReduce/Spark).
- Privacy-preserving mining: anonymization, differential privacy.
- Explainable AI: models must justify decisions in regulated domains.
- Stream mining: patterns from never-ending data (sensors, clicks) in one pass.
- Ethical mining: avoid bias — models trained on biased data discriminate.

Key Points to Remember

- Data mining creates value but also privacy and fairness responsibilities.
- Future = big, streaming, explainable and ethical mining.

Exam Pointer

Very important for exams: Write short notes on any two applications with the technique used — safe 5 marks.

Data mining project pitfalls

- Mining without a business question produces trivia.
- Leaky features inflate accuracy fictitiously.
- Deploy, monitor, maintain — models decay.

Careers in data mining

- Data analyst, data scientist, BI developer, ML engineer.
- SQL + Python + statistics is the entry ticket.
- Pakistani banks, telecoms and startups all hire.

Exam revision checklist

- KDD steps, normalization formulas, Apriori trace.
- Entropy/gain numerical, k-means trace, confusion matrix metrics.
- One comparison table per algorithm pair.

Responsible mining checklist

- Consent, anonymization, bias audit, explainability.
- Document limitations honestly.
- Err on the side of privacy.

Recommendation evaluation

- Precision/recall at k for ranking quality.
- RMSE for rating prediction.
- Diversity and novelty matter to users.
- Offline metrics imperfectly predict online lift.
- A/B test the finalists live.

Data mining pitfalls

- Leakage: future information in training data.
- Overfitting: memorizing noise.
- Wrong metric: optimizing what doesn't matter.
- Ignoring deployment: models that never ship.
- Checklist these before every project.

Glossary of Key Terms

Learning Objectives

- Revise all key terms quickly before exams

Term	Meaning
Data mining	the process of discovering interesting, non-trivial, previously unknown and potentially useful patterns from large amounts of data.
KDD (Knowledge Discovery in Databases)	the overall process of extracting knowledge from data, of which data mining is the central step: selection -> preprocessing -> transformation -> data mining -> ...
Data warehouse	a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management decision-making.
Association rule	an implication of the form $A \rightarrow B$ meaning customers who buy A also tend to buy B, measured by support and confidence.
Classification	supervised learning that builds a model from labelled training data to predict the class of new, unseen tuples.
Clustering	unsupervised grouping of objects so that objects within a cluster are similar and objects across clusters are dissimilar.

Term	Meaning
Outlier	a data object that deviates significantly from the rest — generated by a different mechanism than the normal data.

Exam Pointer

Very important for exams: Revise this glossary the night before the paper.

Practice Questions for Exams

Learning Objectives

- Test your preparation with these exam-style questions

Short questions with model answers

Example

Q: Explain: define data mining and the KDD process Model answer: Data mining = discovery of hidden patterns; KDD = the full pipeline. Know the five functionalities by name.

Example

Q: Explain: describe the kinds of data and patterns mined Model answer: Data mining = discovery of hidden patterns; KDD = the full pipeline. Know the five functionalities by name.

Example

Q: Explain: explain why preprocessing is essential Model answer: Cleaning -> integration -> transformation -> reduction. Normalization is essential before distance-based mining.

Example

Q: Explain: describe cleaning, integration, reduction and transformation Model answer: Cleaning -> integration -> transformation -> reduction. Normalization is essential before distance-based mining.

Example

Q: Explain: define data warehouse and its architecture Model answer: Warehouse = integrated history for analysis; OLTP = live transactions. Star vs snowflake schema difference is a favourite question.

Example

Q: Explain: explain OLAP operations Model answer: Warehouse = integrated history for analysis; OLTP = live transactions. Star vs snowflake schema difference is a favourite question.

Example

Q: Explain: define support and confidence Model answer: Support = frequency; confidence = strength; Apriori prunes with the subset property. FP-growth avoids candidate generation using the FP-tree.

Example

Q: Explain: explain the Apriori algorithm step by step Model answer: Support = frequency; confidence = strength; Apriori prunes with the subset property. FP-growth avoids candidate generation using the FP-tree.

Example

Q: Explain: build a decision tree using information gain Model answer: Decision trees: interpretable, need pruning. Naive Bayes: fast, probabilistic, independence assumption. Overfitting is the main enemy — prune trees and validate.

Example

Q: Explain: explain Naive Bayesian classification Model answer: Decision trees: interpretable, need pruning. Naive Bayes: fast, probabilistic, independence assumption. Overfitting is the main enemy — prune trees and validate.

Example

Q: Explain: explain kNN, neural networks and SVMs Model answer: Accuracy misleads on imbalanced data — use precision/recall/F1. Always evaluate on unseen data via cross-validation.

Example

Q: Explain: describe model evaluation techniques Model answer: Accuracy misleads on imbalanced data — use precision/recall/F1. Always evaluate on unseen data via cross-validation.

Example

Q: Explain: explain k-means and k-medoids Model answer: k-means = fast but needs k and hates outliers; k-medoids = robust but slower. Clustering has no labels — evaluation is harder than classification.

Example

Q: Explain: discuss how to choose the number of clusters Model answer: k-means = fast but needs k and hates outliers; k-medoids = robust but slower. Clustering has no labels — evaluation is harder than classification.

Example

Q: Explain: explain agglomerative vs divisive clustering Model answer: Hierarchical gives a dendrogram; DBSCAN finds arbitrary shapes + outliers. DBSCAN needs eps and minPts tuning.

Example

Q: Explain: describe DBSCAN and its parameters Model answer: Hierarchical gives a dendrogram; DBSCAN finds arbitrary shapes + outliers. DBSCAN needs eps and minPts tuning.

Example

Q: Explain: define outliers and their causes Model answer: Outliers can be errors to remove or the most interesting discoveries (fraud). Always investigate before deleting.

Example

Q: Explain: describe statistical, proximity and clustering-based detection Model answer: Outliers can be errors to remove or the most interesting discoveries (fraud). Always investigate before deleting.

Example

Q: Explain: explain text mining steps Model answer: TF-IDF = term frequency x inverse document frequency. Web mining splits into content, structure and usage.

Example

Q: Explain: describe web mining types and sequential pattern mining Model answer: TF-IDF = term frequency x inverse document frequency. Web mining splits into content, structure and usage.

Example

Q: Explain: list real-world applications Model answer: Data mining creates value but also privacy and fairness responsibilities. Future = big, streaming, explainable and ethical mining.

Example

Q: Explain: discuss privacy, big data and future trends Model answer: Data mining creates value but also privacy and fairness responsibilities. Future = big, streaming, explainable and ethical mining.

Long questions

- Write a detailed note on Introduction to Data Mining.
- Write a detailed note on Data Preprocessing.
- Write a detailed note on Data Warehousing and OLAP.
- Write a detailed note on Mining Frequent Patterns and Association Rules.
- Write a detailed note on Classification: Decision Trees and Bayesian Methods.
- Write a detailed note on Classification: Advanced Methods.
- Write a detailed note on Cluster Analysis: Partitioning Methods.
- Write a detailed note on Cluster Analysis: Hierarchical and Density Methods.
- Write a detailed note on Outlier and Anomaly Detection.
- Write a detailed note on Mining Text, Web and Sequential Data.

Long questions with model answer outlines

Example

Q: Write a detailed note on Introduction to Data Mining. Answer outline: • Why now: data is collected everywhere (transactions, web, sensors) but human analysis cannot keep up. • Multidisciplinary: draws on databases, statistics, machine learning, visualization and high-performance computing. • Data mining vs OLAP: OLAP summarizes and reports; data mining discovers hidden patterns and predicts. • Characterization and discrimination: summarize a target class and contrast it with others. • Frequent pattern mining: association and correlation (market basket analysis).

Example

Q: Write a detailed note on Data Preprocessing. Answer outline: • Missing values: ignore the tuple, fill manually, use the attribute mean/median/mode, or use the most probable value. • Noisy data: binning (smooth by bin means/medians/boundaries), regression smoothing, clustering to detect outliers. • Integration: combine multiple sources; resolve entity identification, redundancy (correlation analysis) and value conflicts. • Normalization: min-max scaling to [0,1], z-score normalization, decimal scaling — needed so large-valued attributes do not dominate distance measures. • Aggregation, generalization and attribute construction create better features.

Example

Q: Write a detailed note on Data Warehousing and OLAP. Answer outline: • OLTP (operational databases): day-to-day transactions, detailed, current. • OLAP (warehouses): analysis, summarized, historical. • Architecture: data sources -> ETL (extract, transform, load) -> warehouse -> data marts -> OLAP tools. • Facts: numeric measures (sales amount). Dimensions: perspectives (time, product, location). • Star schema: one central fact table, denormalized dimension tables.

Example

Q: Write a detailed note on Mining Frequent Patterns and Association Rules. Answer outline: • Apriori property: all non-empty subsets of a frequent itemset must also be frequent — this prunes candidates. • Step 1: find frequent 1-itemsets (count $\geq$ min support). • Step 2: join frequent k-itemsets to form (k+1)-candidates; prune using the Apriori property. • Step 3: scan the database to count candidates; repeat until no more frequent itemsets. • Step 4: generate rules from frequent itemsets and keep those meeting min confidence.

Example

Q: Write a detailed note on Classification: Decision Trees and Bayesian Methods. Answer outline: • Each internal node tests an attribute, branches show outcomes, leaves give class labels. • Attribute selection: choose the attribute with highest information gain (or gain ratio, or Gini index). • Entropy measures impurity; information gain = reduction in entropy after splitting. • Tree pruning (prepruning/postpruning) removes overfitting branches. • Based on Bayes theorem: $P(H|X) = P(X|H) \times P(H) / P(X)$.

Example

Q: Write a detailed note on Classification: Advanced Methods. Answer outline: • k-Nearest Neighbours: lazy learner — stores all training data, classifies by majority vote of k nearest neighbours using a distance metric. Simple, no training, but slow at prediction. • Neural networks: layers of neurons with weights learned by backpropagation; powerful for complex patterns but "black box" and slow to train. • Support Vector Machines: find the maximum-margin hyperplane separating classes; excellent for high-dimensional data like text. • Confusion matrix organizes TP, TN, FP, FN. • Holdout, cross-validation (k-fold) and bootstrap estimate true performance.

Example

Q: Write a detailed note on Cluster Analysis: Partitioning Methods. Answer outline: • 1. Choose k initial centroids. 2. Assign each point to the nearest centroid. 3. Recompute centroids as cluster means. 4. Repeat until stable. • Strengths: simple, efficient, scalable. Weaknesses: must specify k, sensitive to outliers and initial centroids, works poorly on non-spherical clusters. • k-medoids (PAM) uses actual data points as centres — more robust to outliers but costlier. • Elbow method: plot within-cluster sum of squares vs k; look for the "elbow". • Silhouette coefficient measures how well each point fits its cluster (-1 to 1).

Example

Q: Write a detailed note on Cluster Analysis: Hierarchical and Density Methods. Answer outline: • Agglomerative (bottom-up): each point starts as its own cluster; merge the closest pair repeatedly until one cluster remains. • Divisive (top-down): start with one cluster and split. • Linkage criteria: single (nearest points), complete (farthest), average, centroid. • Result shown as a dendrogram; cut it at a chosen height to get clusters. • Parameters: eps (neighbourhood radius) and minPts (minimum points to form a dense region).

Example

Q: Write a detailed note on Outlier and Anomaly Detection. Answer outline: • Causes: measurement errors, fraud, intrusions, rare events, or data entry mistakes. • Applications: credit-card fraud, network intrusion, faulty sensors, medical anomalies. • Statistical: model normal data with a distribution; flag points with low probability (z-score > 3). • Proximity-based: points far from their k nearest neighbours are outliers. • Clustering-based: small clusters or points far from any cluster centre are suspicious.

Example

Q: Write a detailed note on Mining Text, Web and Sequential Data. Answer outline: • Pipeline: tokenization -> stop-word removal -> stemming/lemmatization -> vectorization (bag of words, TF-IDF). • TF-IDF weights terms by frequency in the document vs rarity across documents. • Applications: sentiment analysis, spam filtering, topic modelling. • Content mining: extract knowledge from page text. • Structure mining: analyse hyperlinks (PageRank ranks pages by link importance).

Exam Pointer

Very important for exams: Practice writing answers within the exam time limit.

MCQ Practice Test

Learning Objectives

- Test your recall of the key points from every chapter

Q1. Which of the following was listed as a key point of 'Introduction to Data Mining' in this handout?

- A. Data mining = discovery of hidden patterns; KDD = the full pipeline.
- B. Warehouse = integrated history for analysis; OLTP = live transactions.
- C. Normalization is essential before distance-based mining.
- D. Cleaning -> integration -> transformation -> reduction.

Q2. Which of the following was listed as a key point of 'Introduction to Data Mining' in this handout?

- A. Warehouse = integrated history for analysis; OLTP = live transactions.
- B. Know the five functionalities by name.
- C. Normalization is essential before distance-based mining.
- D. Cleaning -> integration -> transformation -> reduction.

Q3. Which of the following was listed as a key point of 'Data Preprocessing' in this handout?

- A. Cleaning -> integration -> transformation -> reduction.
- B. Know the five functionalities by name.
- C. Warehouse = integrated history for analysis; OLTP = live transactions.
- D. Data mining = discovery of hidden patterns; KDD = the full pipeline.

Q4. Which of the following was listed as a key point of 'Data Preprocessing' in this handout?

- A. Know the five functionalities by name.
- B. Normalization is essential before distance-based mining.
- C. Warehouse = integrated history for analysis; OLTP = live transactions.
- D. Data mining = discovery of hidden patterns; KDD = the full pipeline.

Q5. Which of the following was listed as a key point of 'Data Warehousing and OLAP' in this handout?

- A. Cleaning -> integration -> transformation -> reduction.

B. Warehouse = integrated history for analysis; OLTP = live transactions.

C. Data mining = discovery of hidden patterns; KDD = the full pipeline.

D. Know the five functionalities by name.

Q6. Which of the following was listed as a key point of 'Data Warehousing and OLAP' in this handout? 

A. Cleaning -> integration -> transformation -> reduction.

B. Know the five functionalities by name.

C. Star vs snowflake schema difference is a favourite question.

D. Data mining = discovery of hidden patterns; KDD = the full pipeline.

Q7. Which of the following was listed as a key point of 'Mining Frequent Patterns and Association Rules' in this handout?

A. Support = frequency; confidence = strength; Apriori prunes with the subset property.

B. Data mining = discovery of hidden patterns; KDD = the full pipeline.

C. Know the five functionalities by name.

D. Cleaning -> integration -> transformation -> reduction.

Q8. Which of the following was listed as a key point of 'Mining Frequent Patterns and Association Rules' in this handout?

A. Data mining = discovery of hidden patterns; KDD = the full pipeline.

B. Know the five functionalities by name.

C. Cleaning -> integration -> transformation -> reduction.

D. FP-growth avoids candidate generation using the FP-tree.

Q9. Which of the following was listed as a key point of 'Classification: Decision Trees and Bayesian Methods' in this handout?

A. Cleaning -> integration -> transformation -> reduction.

B. Know the five functionalities by name.

C. Decision trees: interpretable, need pruning. Naive Bayes: fast, probabilistic, independence assumption.

D. Data mining = discovery of hidden patterns; KDD = the full pipeline.

Q10. Which of the following was listed as a key point of 'Classification: Decision Trees and Bayesian Methods' in this handout?

A. Overfitting is the main enemy — prune trees and validate.

B. Know the five functionalities by name.

C. Cleaning -> integration -> transformation -> reduction.

D. Data mining = discovery of hidden patterns; KDD = the full pipeline.

Q11. Which of the following was listed as a key point of 'Classification: Advanced Methods' in this handout? 

A. Data mining = discovery of hidden patterns; KDD = the full pipeline.

B. Know the five functionalities by name.

C. Cleaning -> integration -> transformation -> reduction.

D. Accuracy misleads on imbalanced data — use precision/recall/F1.

Q12. Which of the following was listed as a key point of 'Classification: Advanced Methods' in this handout?

A. Always evaluate on unseen data via cross-validation.

B. Data mining = discovery of hidden patterns; KDD = the full pipeline.

C. Cleaning -> integration -> transformation -> reduction.

D. Know the five functionalities by name.

Q13. Which of the following was listed as a key point of 'Cluster Analysis: Partitioning Methods' in this handout?

A. Cleaning -> integration -> transformation -> reduction.

B. Data mining = discovery of hidden patterns; KDD = the full pipeline.

C. k-means = fast but needs k and hates outliers; k-medoids = robust but slower.

D. Know the five functionalities by name.

Q14. Which of the following was listed as a key point of 'Cluster Analysis: Partitioning Methods' in this handout?

A. Know the five functionalities by name.

B. Cleaning -> integration -> transformation -> reduction.

C. Clustering has no labels — evaluation is harder than classification.

D. Data mining = discovery of hidden patterns; KDD = the full pipeline.

Q15. Which of the following was listed as a key point of 'Cluster Analysis: Hierarchical and Density Methods' in this handout?

A. Know the five functionalities by name.

B. Hierarchical gives a dendrogram; DBSCAN finds arbitrary shapes + outliers.

C. Cleaning -> integration -> transformation -> reduction.

D. Data mining = discovery of hidden patterns; KDD = the full pipeline.

Q16. Which of the following was listed as a key point of 'Cluster Analysis: Hierarchical and Density Methods' in this handout?

- A. DBSCAN needs eps and minPts tuning.
- B. Cleaning -> integration -> transformation -> reduction.
- C. Know the five functionalities by name.
- D. Data mining = discovery of hidden patterns; KDD = the full pipeline.

Q17. Which of the following was listed as a key point of 'Outlier and Anomaly Detection' in this handout? 

- A. Data mining = discovery of hidden patterns; KDD = the full pipeline.
- B. Outliers can be errors to remove or the most interesting discoveries (fraud).
- C. Cleaning -> integration -> transformation -> reduction.
- D. Know the five functionalities by name.

Q18. Which of the following was listed as a key point of 'Outlier and Anomaly Detection' in this handout?

- A. Data mining = discovery of hidden patterns; KDD = the full pipeline.
- B. Cleaning -> integration -> transformation -> reduction.
- C. Know the five functionalities by name.
- D. Always investigate before deleting.

Q19. Which of the following was listed as a key point of 'Mining Text, Web and Sequential Data' in this handout?

- A. Data mining = discovery of hidden patterns; KDD = the full pipeline.
- B. Cleaning -> integration -> transformation -> reduction.
- C. TF-IDF = term frequency x inverse document frequency.
- D. Know the five functionalities by name.

Q20. Which of the following was listed as a key point of 'Mining Text, Web and Sequential Data' in this handout?

- A. Data mining = discovery of hidden patterns; KDD = the full pipeline.
- B. Know the five functionalities by name.
- C. Cleaning -> integration -> transformation -> reduction.
- D. Web mining splits into content, structure and usage.

Q21. Which of the following was listed as a key point of 'Data Mining Applications and Trends' in this handout?

- A. Data mining creates value but also privacy and fairness responsibilities.
- B. Know the five functionalities by name.
- C. Data mining = discovery of hidden patterns; KDD = the full pipeline.

D. Cleaning -> integration -> transformation -> reduction.

Q22. Which of the following was listed as a key point of 'Data Mining Applications and Trends' in this handout? 

A. Cleaning -> integration -> transformation -> reduction.

B. Data mining = discovery of hidden patterns; KDD = the full pipeline.

C. Know the five functionalities by name.

D. Future = big, streaming, explainable and ethical mining.

Answer Key

1. A 2. B 3. A 4. B 5. B 6. C 7. A 8. D 9. C 10. A 11. D 12. A 13. C 14. C 15. B 16. A 17. B 18. D 19. C 20. D 21. A 22. D

Exam Pointer

Very important for exams: Attempt all MCQs without looking back, then check the answer key.

Exam Strategy Guide

Learning Objectives

- Attempt papers with a winning strategy

Before the exam

- Revise the Key Points boxes and glossary first — they carry the highest marks per minute.
- Practice one past paper under timed conditions; examiners repeat question styles.
- Sleep well the night before — an exhausted brain recalls far less.

Attempting the paper

- Read the whole paper first; start with your strongest question to build momentum.
- Allocate time by marks: a 10-mark question deserves roughly double a 5-mark one.
- Structure long answers: definition, explanation, example/diagram, conclusion.

Presentation matters

- Use headings, underline key terms, and draw labelled diagrams where possible.
- Attempt all questions — partial credit beats a blank page.
- Review your paper in the last 10 minutes for missed parts.

Exam Pointer

Very important for exams: Very important for exams: neat, structured answers consistently score higher than longer messy ones.