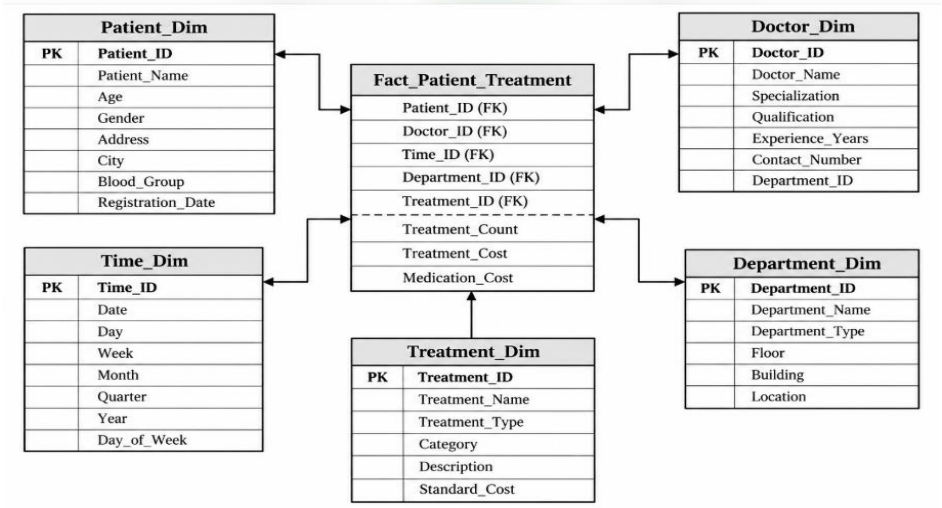
| (Affiliated to University of Mumbai) | ***End Semester Examination (R-24/R-25) SH 2026 Answer Key with marking scheme*** |
|---|---|
| Branch: AIDS | Course: Data Warehousing and Mining |
| Year/ Semester: 2026/V | Course code: AIDSC503 |
| Time: 03 hours | Marks: 80 |

| | | Marks |
|---|---|---|
| Q. 1 | All questions are compulsory. | |
| A. | OLAP Cube-2M, multidimensional analysis-3M<br>An OLAP cube is a multidimensional data structure used to organize and analyze data across dimensions such as time, product, and location. For example, a sales OLAP cube helps analyze total sales by product, month, and region using operations such as slice, dice, and roll-up. | 5 |
| B. | Types-2M, proximity measure-3M<br>Attributes in data exploration include nominal, binary, ordinal, and numeric attributes (interval and ratio). Proximity measures for binary attributes include the Jaccard coefficient for asymmetric binary data | 5 |
| C. | 2M-for each calculation<br>Mean $= \frac{203}{10}$ =20.3, Mode=18, midrange=(min+max)/2=20.5, variance<br>$\sigma^2 = \frac{\sum x^2}{n} - (\bar{x})^2$; ( x^2=4501) and {x}=20.3)=38.01 | 5 |
| D. | Definitiob-1M, step1-2M, step2-2M<br>Clustering is an unsupervised machine learning technique that groups similar data points into clusters based on their characteristics or distance, such that points within the same cluster are more similar to each other than to those in other clusters [2]<br><br>Assuming initial centroids (C_1=3\) and (C_2=27\), the iterations converge to centroids (C_1=9.2) and (C_2=21). Final clusters: (C_1={3,6,8,14,16}), (C_2={19,23,27}); clustering groups similar data points based on distance , [3] | 5 |
| E. | Idea-2M, advantages-3M<br>DBSCAN groups closely packed data points into clusters using two parameters: Epsilon (ε) and MinPts. It identifies core, border, and noise points based on density and effectively detects outliers without requiring the number of clusters in advance. | 5 |
| F. | Explanation-3M, example-2M<br>Multilevel association rule mining discovers relationships among items at different levels of abstraction using a concept hierarchy. For example, a rule may show that customers who buy dairy products also buy bread, while a more specific rule may show that customers who buy milk also buy brown bread. | 5 |

| | | |
|---|---|---|
| Q.2A | <br>[5]<br>Given: Time period = 5 years, Departments = 10,Patients = 50,000,Doctors = 500,Maximum treatment records per patient per day = 2<br>Assuming 365 days per year:<br>Total days=5X365=1825<br>Since each patient can have at most 2 treatment records per day:1825X50000X2=182500000 [5] | 10 |
| OR | | |
| Q.2B. | ETL Process Overview: [2]<br>Extraction of data from multiple sources: [2]<br>Transformation (cleaning, integration, and formatting): [3]<br>Loading data into the warehouse with a suitable diagram/example: [3]<br>Data Sources<br>↓<br>Extract → Transform → Load<br>↓<br>Data Warehouse | 10 |
| | | |
| Q.3A | 2M -each<br>TP=60,FN=20,FP=10,TN=110<br><br>**Metric** / **Formula** / **Calculation and Answer**<br>Accuracy / ((TP + TN) / Total) × 100 / ((60 + 110) / 200) × 100 = 85%<br>Error Rate / ((FP + FN) / Total) × 100 / ((10 + 20) / 200) × 100 = 15%<br>Precision / (TP / (TP + FP)) × 100 / (60 / (60 + 10)) × 100 = 85.71%<br>Recall (Sensitivity) / (TP / (TP + FN)) × 100 / (60 / (60 + 20)) × 100 = 75%<br>F1-score / (2TP / (2TP + FP + FN)) × 100 / (120 / (120 + 10 + 20)) × 100 = 80% | 10 |
| OR | | |
| Q.3B | P(Flu = Y) = 5/8 = 0.625<br>P(Flu = N) = 3/8 = 0.375 | 10 |

| | | |
|---|---|---|
| | | |

| Attribute | Given Flu = Y | Given Flu = N |
|---|---|---|
| Chills = Y | 3/5 | 1/3 |
| Runny nose = N | 1/5 | 2/3 |
| Headache = Mild | 2/5 | 1/3 |
| Fever = Y | 4/5 | 1/3 |

For Flu = Y:

$P(X \mid Y)P(Y) = (5/8) \times (3/5) \times (1/5) \times (2/5) \times (4/5) = 0.024$

For Flu = N:

$P(X \mid N)P(N) = (3/8) \times (1/3) \times (2/3) \times (1/3) \times (1/3) \approx 0.00926$

Since $0.024 > 0.00926$, the Naïve Bayes classifier predicts Flu = Y. Therefore, Patient X is classified as having flu.

| | | |
|---|---|---|
| | | |
| Q.4A | 2M-diagram, explanation -6M<br>1) Data Cleaning: Removes noise, duplicate records, and missing or inconsistent values.<br>2) Data Integration: Combines data from different sources into a unified dataset.<br>3) Data Selection: Selects data relevant to the analysis objective.<br>4) Data Transformation: Converts data into a suitable format using normalization or aggregation.<br>5) Data Mining: Applies algorithms to discover patterns, relationships, or trends.<br>6) Pattern Evaluation: Evaluates patterns to identify useful and meaningful knowledge.<br>7) Knowledge Presentation: Displays the discovered knowledge through reports, charts, or visualizations. | 10 |
| | OR | |
| Q.4B | Concept-2M, 2M-each technique<br>Data preprocessing improves data quality through cleaning, integration, transformation, and reduction.<br>Discretization converts continuous values into intervals, while normalization scales values to a common range.<br>Concept hierarchy generation organizes data into multiple levels of abstraction, such as city → state → country.<br>These techniques reduce complexity, improve data consistency, and increase the efficiency of data mining. | 10 |
| | | |
| Q.5A | MBA-4M, 3M-support, 3M-confidence<br>Market Basket Analysis identifies items frequently purchased together; for example, customers who buy bread may also buy butter.<br>i) Support: Measures how frequently items X and Y occur together.<br>Formula: Support $(X \rightarrow Y)$ = Number of transactions containing both X and Y / Total number of transactions.<br>ii) Confidence: Measures how often Y is purchased when X is purchased. Formula:<br>Confidence$(X \rightarrow Y)$ = [Support$(X \cup Y)$ / Support$(X)$] × 100%.<br>These measures help retailers discover purchasing patterns and make product placement and promotional decisions. | 10 |
| | OR | |

Q.5B

| Itemset | Support Count | Support (%) |
|---|---|---|
| {M1} | 4 | 66.67% |
| {M2} | 5 | 83.33% |
| {M3} | 4 | 66.67% |
| {M4} | 4 | 66.67% |
| {M5} | 2 | 33.33% — not frequent |

| Itemset | Support Count | Support (%) |
|---|---|---|
| {M1, M2} | 4 | 66.67% |
| {M1, M3} | 3 | 50.00% |
| {M2, M3} | 4 | 66.67% |
| {M2, M4} | 3 | 50.00% |

{M1, M2, M3} occurs in T1, T5, and T6. Support count = 3; support = 3/6 × 100 = 50%. Therefore, it is frequent. No larger frequent itemset can be generated.

Confidence(X → Y) = Support Count(X ∪ Y) / Support Count(X) × 100%.

| Rule | Confidence Calculation | Confidence |
|---|---|---|
| M1 → M2 | 4/4 × 100 | 100% |
| M2 → M1 | 4/5 × 100 | 80% |
| M1 → M3 | 3/4 × 100 | 75% |
| M3 → M1 | 3/4 × 100 | 75% |
| M2 → M3 | 4/5 × 100 | 80% |
| M3 → M2 | 4/4 × 100 | 100% |
| M2 → M4 | 3/5 × 100 | 60% |
| M4 → M2 | 3/4 × 100 | 75% |
| M1 → {M2, M3} | 3/4 × 100 | 75% |
| M2 → {M1, M3} | 3/5 × 100 | 60% |
| M3 → {M1, M2} | 3/4 × 100 | 75% |
| {M1, M2} → M3 | 3/4 × 100 | 75% |
| {M1, M3} → M2 | 3/3 × 100 | 100% |
| {M2, M3} → M1 | 3/4 × 100 | 75% |

Frequent itemsets: L1 = {M1}, {M2}, {M3}, {M4}; L2 = {M1,M2}, {M1,M3}, {M2,M3}, {M2,M4}; L3 = {M1,M2,M3}.

Strong association rules are the rules listed above, as each has confidence ≥ 60%. The minimum support threshold is 50% (support count ≥ 3).

Q.6A

5M-HITS, 5M-text generation

The HITS (Hyperlink-Induced Topic Search) algorithm ranks web pages using hub values (pages linking to useful resources) and authority values (pages containing valuable information). It iteratively updates and normalizes these scores until they converge. [5]

The major steps are text collection, preprocessing (tokenization, stop-word removal, stemming), feature extraction, classifier training, and classification/evaluation. For example, an email system categorizes messages as spam or not spam using a trained machine learning model [5]

OR

Q.6B

5M-crawlers, 5M-personalization

10

Web Crawler Architecture: A web crawler consists of a URL frontier, crawler, page downloader, parser, and repository. It fetches web pages, extracts hyperlinks, adds new URLs to the frontier, and stores collected content for analysis.

2. Working: The crawler starts with seed URLs, downloads pages, extracts relevant content and links, and repeats the process while avoiding duplicate URLs. [6]

3. Personalization: Personalization uses user preferences, browsing history, and interests to customize content and recommendations. [4]

*********** All the Best***********