VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

**DMT(22CS306)-MODULE BANK**

**Department of CSE**

## 1) Weather Station Data Analysis

**Dataset**:

| Station | Temp (°C) | Humidity (%) | Wind Speed (km/h) | AQI |
|---|---|---|---|---|
| S1 | 32 | 45 | 20 | 100 |
| S2 | 30 | 50 | 18 | 90 |
| S3 | 34 | 40 | 24 | 120 |
| S4 | 28 | 55 | 16 | 80 |
| S5 | 31 | 48 | 22 | 110 |
| S6 | 33 | 42 | 20 | 105 |
| S7 | 29 | 60 | 15 | 95 |
| S8 | 35 | 38 | 26 | 130 |
| S9 | 30 | 52 | 17 | 92 |
| S10 | 27 | 65 | 12 | 85 |

**a)** Define and classify each attribute into nominal, ordinal, interval, or ratio types. For example, the station name is nominal, temperature and AQI are ratios, and so on. Explain how understanding the attribute types guides preprocessing decisions such as normalization, encoding, or discretization techniques.

**b)** Describe how central tendency metrics like mean and median help summarize the dataset. Similarly, explain how dispersion metrics such as standard deviation and range can highlight variability in the data. Illustrate how AQI values can be interpreted using these metrics to identify air quality patterns across stations.

**c)**Manually calculate the mean temperature and AQI across all the stations. Interpret whether the average values fall within acceptable environmental standards or indicate pollution risks. Provide insights on what this means for public health or climate conditions.

**d)**Use the dataset to examine if a pattern exists between temperature and AQI do higher temperatures lead to higher AQI? Frame a hypothesis, such as "Increasing temperatures are correlated with declining air quality." Discuss how this hypothesis could be tested using correlation or regression.

**2) Student Performance and Behaviour Analysis**

**Dataset**:

| ID | Study Hours | Attendance (%) | Social Media (hrs/day) | Score |
|---|---|---|---|---|
| S1 | 5 | 92 | 1.5 | 88 |
| S2 | 3 | 85 | 4.0 | 70 |
| S3 | 6 | 95 | 1.0 | 90 |
| S4 | 2 | 80 | 3.5 | 65 |
| S5 | 7 | 98 | 1.2 | 92 |
| S6 | 4 | 75 | 2.5 | 72 |
| S7 | 5 | 88 | 2.0 | 84 |
| S8 | 6 | 90 | 1.0 | 89 |
| S9 | 3 | 78 | 4.2 | 68 |
| S10 | 2 | 70 | 5.0 | 55 |

**a)** Identify all attributes and classify them based on their types (e.g., ratio, nominal). Explain why recognizing numerical vs. categorical features matters for tasks like normalization or encoding. Highlight how different model types handle these attributes during training.

**b)** Discuss how social media hours may negatively or positively correlate with academic performance. Support your reasoning with examples from the dataset or hypothetical scenarios. Suggest how statistical techniques or visualizations can confirm this relationship.

**c)**Demonstrate min-max normalization for “Score” and “Study Hours” for three students. Clearly show the transformation formula and calculations. Explain the purpose of scaling in ML pipelines, especially for distance-based algorithms.

**d)**Compare the importance of features such as study hours, attendance, and social media use in predicting scores. Argue whether a linear model would suffice or if a non-linear tree-based model could capture more complex interactions. Justify your choice based on interpretability and accuracy.

**3) Customer Spending Patterns**

**Dataset**:

| ID | Electronics | Groceries | Clothing | Home Decor | Total Bill ($) |
|---|---|---|---|---|---|
| C1 | 120 | 75 | 80 | 45 | 320 |
| C2 | 200 | 60 | 100 | 80 | 440 |
| C3 | 50 | 90 | 70 | 20 | 230 |
| C4 | 150 | 40 | 60 | 30 | 280 |
| C5 | 180 | 85 | 110 | 90 | 465 |
| C6 | 130 | 70 | 90 | 60 | 350 |

| ID | Electronics | Groceries | Clothing | Home Decor | Total Bill ($) |
|---|---|---|---|---|---|
| C7 | 160 | 50 | 95 | 35 | 340 |
| C8 | 90 | 95 | 60 | 55 | 300 |
| C9 | 170 | 45 | 100 | 40 | 355 |
| C10 | 60 | 85 | 55 | 25 | 225 |

**a)** List and describe numeric attributes like Electronics, Groceries, etc., and how they represent spending behaviour. Explain how these metrics assist in building customer profiles. Discuss the potential role of each attribute in segmenting and targeting.

**b)** Analyze how differing amounts spent in categories reflect customer preferences or lifestyle. For instance, a customer spending more on groceries may represent a budget-conscious household. Discuss how segmentation helps in tailoring personalized marketing campaigns.

**c)**Perform a correlation analysis between Electronics and Total Bill attributes. Interpret whether there is a strong, weak, or no correlation. Discuss the implication for bundling or recommending electronics to high-spending customers.

**d)**Explain how clustering techniques like k-means can group customers with similar spending habits. Describe the benefits of customer segmentation for designing targeted offers. Mention possible challenges such as feature scaling or number of clusters.

**4) Online Course Completion Analysis**

**Dataset**:

| User | Videos Watched (%) | Quizzes Taken | Time Spent (hrs) | Completed |
|---|---|---|---|---|
| U1 | 90 | 10 | 15 | Yes |
| U2 | 40 | 3 | 6 | No |
| U3 | 75 | 8 | 10 | Yes |
| U4 | 35 | 2 | 5 | No |
| U5 | 95 | 9 | 18 | Yes |
| U6 | 20 | 1 | 4 | No |
| U7 | 80 | 7 | 13 | Yes |
| U8 | 60 | 5 | 9 | Yes |
| U9 | 45 | 3 | 7 | No |
| U10 | 100 | 10 | 20 | Yes |

**a)** Identify predictors like videos watched, quizzes taken, time spent, and the target variable (completion status). Explain how each feature contributes to predicting completion likelihood. Discuss the importance of binary classification in e-learning systems.

**b)** Describe how decision trees split data based on features like quizzes and videos to predict outcomes. Explain how the tree learns from data using entropy or Gini index. Emphasize its interpretability and relevance in educational analytics.

**c)**Create a sample rule such as: “If quizzes taken ≥ 5 and videos watched ≥ 70%, then the student is likely to complete the course.” Justify the logic using the dataset. Discuss how such rules aid instructors in identifying at-risk students.

**d)**Discuss the role of “time spent” in measuring student engagement. Analyze if it could be misleading e.g., idle tab time might inflate time but not reflect learning. Provide recommendations on how to accurately capture engagement.

**5) Loan Risk Categorization**

**Dataset**:

| ID | Age | Income (k) | Credit Score | Loan Amount (k) | Defaulted |
|---|---|---|---|---|---|
| L1 | 25 | 30 | 680 | 10 | No |
| L2 | 45 | 60 | 720 | 20 | No |
| L3 | 35 | 50 | 650 | 15 | Yes |
| L4 | 29 | 40 | 630 | 12 | Yes |
| L5 | 50 | 80 | 700 | 25 | No |
| L6 | 42 | 55 | 610 | 18 | Yes |
| L7 | 33 | 52 | 640 | 14 | Yes |
| L8 | 28 | 45 | 670 | 13 | No |
| L9 | 47 | 70 | 690 | 22 | No |
| L10 | 38 | 48 | 600 | 15 | Yes |

**a)** List the attributes and identify which are predictors (e.g., age, income) and which is the target (defaulted). Explain why this classification is important for supervised learning tasks. Describe how these attributes influence model training.

**b)** Describe how feature selection eliminates redundant or irrelevant attributes to improve model performance. Emphasize its role in reducing overfitting and speeding up training. Give examples of filters like variance threshold or mutual information.

**c)**Suggest a strategy to discretize income into bins: e.g., Low (<45k), Medium (45–65k), High (>65k). Explain how discretization helps algorithms that work better with categorical inputs. Discuss trade-offs such as loss of granularity.

**d)**Analyze which features (like credit score or loan amount) have a strong impact on defaulting. Use correlation reasoning or domain knowledge to support your claims. Justify why certain features are prioritized in credit risk models.

**6) Weekly Sales Record for Categories**

**Dataset**:

| ID | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday | Category |
|---|---|---|---|---|---|---|---|---|
| P1 | 50 | 40 | 30 | 35 | 60 | 70 | 90 | Electronics |
| P2 | 80 | 70 | 65 | 60 | 75 | 85 | 100 | Groceries |
| P3 | 40 | 30 | 25 | 35 | 55 | 65 | 80 | Home Goods |
| P4 | 100 | 90 | 85 | 80 | 95 | 110 | 120 | Electronics |
| P5 | 30 | 25 | 20 | 22 | 35 | 50 | 70 | Stationery |
| P6 | 90 | 80 | 70 | 75 | 85 | 95 | 110 | Electronics |
| P7 | 60 | 50 | 45 | 48 | 60 | 75 | 95 | Groceries |
| P8 | 20 | 15 | 10 | 12 | 18 | 25 | 30 | Stationery |
| P9 | 55 | 45 | 40 | 42 | 50 | 60 | 75 | Home Goods |
| P10 | 70 | 60 | 55 | 58 | 65 | 80 | 90 | Groceries |

**a)** List daily sales as time-series features and explain their importance in sales trend analysis. Discuss how variations across the week indicate consumer habits. Emphasize the need for time-based features in forecasting.

**b)** Explain how plotting weekly trends for each product helps identify high-performing days or dips. Use example products like Electronics to show pattern recognition. Suggest visualizations like line plots or heatmaps.

**c)**Apply z-score normalization on Monday sales data. Demonstrate the formula and calculations. Explain how standardization allows comparisons across different sales categories.

**d)**Discuss how low sales over the week can indicate underperforming categories. Suggest a ranking approach based on total weekly sales or average daily sales. Explain how such analysis guides promotional planning.

**7) Hospital Patient Records – Data Cleaning & Integration**

**Dataset**:

| Patient ID | Name | Age | Gender | BP (mmHg) | Email |
|---|---|---|---|---|---|
| P1 | Rakesh | 35 | M | 120/80 | rakesh@abc.com |
| P2 | Anjali | | F | 130/90 | anjali@abc.com |
| P3 | Pradeep | 28 | M | 115/75 | pradeep123 |
| P4 | Rekha | 43 | F | | rekha@abc.com |
| P5 | Sameer | 37 | M | 122/84 | sameer@abc.com |
| P6 | | 30 | M | 125/80 | unknown |
| P7 | Harsha | 29 | | 118/78 | harsha@abc.com |
| P8 | Anita | 31 | F | 135/95 | anita@abc.com |

| Patient ID | Name | Age | Gender | BP (mmHg) | Email |
|---|---|---|---|---|---|
| P9 | Sunil | 41 | M | 120/85 | sunil@abc.com |
| P10 | Lavanya | 26 | F | 112/70 | lavanya@abc.com |

**a)** Define common issues like missing values (e.g., Age), inconsistent entries (e.g., Email), and noise (e.g., wrong formats). Identify examples from the dataset. Discuss how these issues hinder accurate analysis and lead to faulty models.

**b)** Describe data cleaning methods like mean imputation for missing Age, regex correction for email, or mode filling for gender. Emphasize that high-quality health records are critical for diagnosis, reporting, and analytics. Highlight the risk of erroneous decisions due to unclean data.

**c)**Apply cleaning steps for five records: fill missing ages, format incorrect emails, and resolve unknown genders. Provide a before-and-after table. Explain how these cleaned records now enable valid model training.

**d)**Explain how classification tasks like identifying high-BP patients fail when data is missing or incorrect. Discuss how model accuracy improves after cleaning. Support with reasoning on interpretability and decision confidence.

**8) Fitness Tracker Dataset – Transformation and Discretization**

**Dataset**:

| User ID | Daily Steps | Calories Burned | Sleep Hours | Active Minutes |
|---|---|---|---|---|
| U1 | 9500 | 320 | 7.5 | 40 |
| U2 | 4800 | 200 | 6.0 | 25 |
| U3 | 11000 | 380 | 8.0 | 50 |
| U4 | 7500 | 290 | 7.0 | 35 |
| U5 | 3000 | 150 | 5.5 | 20 |
| U6 | 8200 | 310 | 6.5 | 38 |
| U7 | 4200 | 190 | 5.0 | 22 |
| U8 | 9800 | 340 | 7.2 | 42 |
| U9 | 5600 | 240 | 6.0 | 30 |
| U10 | 10300 | 360 | 7.8 | 45 |

**a)** Define normalization (rescaling values) and discretization (converting numerical into categories). Explain why these steps are vital for clustering and classification. Mention that inconsistent scales can mislead distance-based models.

**b)** Discuss how z-score keeps mean zero and unit variance, while min-max scales to [0,1]. Compare how “Daily Steps” with high range differs from “Sleep Hours” with low range. Show why uniform scaling ensures fair weightage.

**c)**Perform min-max normalization for "Daily Steps" and bin "Sleep Hours" into three categories: Low (≤6), Medium (6–7.5), High (>7.5). Justify how discretization helps simplify and interpret user behavior. Use at least three samples.

**d)**Explain how transformation helps detect outliers, e.g., unusually high steps. Discuss how binning simplifies downstream analysis like clustering health groups. Mention the trade-off between interpretability and precision.

**9) Sales Dataset – Attribute Selection**

**Dataset**:

| Product ID | Price | Discount (%) | Stock | Views | Sales Count |
|---|---|---|---|---|---|
| P1 | 500 | 10 | 120 | 800 | 90 |
| P2 | 750 | 15 | 100 | 1200 | 130 |
| P3 | 300 | 5 | 80 | 600 | 50 |
| P4 | 650 | 20 | 70 | 1000 | 110 |
| P5 | 400 | 0 | 90 | 700 | 65 |
| P6 | 850 | 25 | 60 | 1500 | 140 |
| P7 | 700 | 15 | 110 | 1100 | 125 |
| P8 | 450 | 8 | 100 | 850 | 80 |
| P9 | 600 | 12 | 95 | 900 | 100 |
| P10 | 520 | 10 | 105 | 950 | 95 |

**a)** Define attribute selection as choosing the most relevant features from a dataset. Explain how removing low-impact features improves model efficiency and prevents overfitting. Provide examples of irrelevant vs. high-value attributes.

**b)** Differentiate filter methods (e.g., correlation), wrapper (model-based selection), and embedded (e.g., Lasso). Recommend which suits a retail dataset where attributes are numeric and interpretable. Justify with pros and cons.

**c)**Calculate the correlation between "Views" and "Sales Count," and between "Discount" and "Sales Count." Compare correlation coefficients to determine the stronger predictor. Interpret how these results impact pricing strategies.

**d)**Discuss how including all features can over fit the model to noise or rare patterns. Highlight the importance of balancing model accuracy and simplicity. Describe real-time system challenges like latency and model interpretability.

**10) Supermarket Transactions – Association Rule Mining**

**Dataset (Transactions)**:

| TID | Items Bought |
|---|---|
| 1 | Milk, Bread, Eggs |

| TID | Items Bought |
|---|---|
| 2 | Milk, Butter, Eggs |
| 3 | Bread, Butter |
| 4 | Milk, Bread, Eggs, Butter |
| 5 | Eggs, Coke |
| 6 | Bread, Coke |
| 7 | Milk, Eggs, Coke |
| 8 | Bread, Butter, Eggs |
| 9 | Milk, Bread |
| 10 | Milk, Butter |

**a)**Define support (frequency), confidence (accuracy), and lift (strength) for association rules. Explain their significance in evaluating rule usefulness. Provide examples from the dataset to illustrate their computation.

**b)**Explain how Apriori generates 1-itemsets, then builds larger itemsets by joining and pruning. Use a dataset transaction to walk through the generation. Emphasize how pruning improves efficiency by removing unlikely combinations.

**c)**Generate all frequent 2-itemsets with minimum support of 0.3 (i.e., in at least 3 out of 10 transactions). Then compute confidence and lift for rule {Milk} → {Eggs}. Interpret the results and their meaning for retail.

**d)**Identify the strongest rule from the dataset and discuss its business impact. For example, a strong lift in {Bread, Butter} may indicate bundling opportunity. Suggest how the store can optimize product placement and promotions.

.