# Question 1

#### Problem Statement

You are a data analyst at a popular e-commerce company. The company has observed that there are often patterns in the way customers buy products. They are interested in uncovering these patterns to make better product recommendations. The dataset provided to you contains a list of transactions, where each transaction is a list of products purchased by a customer.

## Part A: Theoretical Aspects

1. Define the concepts of support, confidence, and lift in the context of association rule mining.

2. Explain how the Apriori algorithm works. Why is it important to prune the itemsets in the process of finding frequent itemsets?

3. Given the following transactions:

A list of food items

Description automatically generated

Calculate the support for the itemset {Bread, Milk} and the confidence for the rule Bread -> Milk. Assume the minimum support threshold is 0.2 and the minimum confidence threshold is 0.5.

## Part B: Implementation

Below is a Python code snippet that generates a list of transactions. Each transaction is represented as a list of products.

```python

products = ["Laptop", "Phone", "Headphones", "Charger", "Mouse", "Keyboard", "Monitor", "Tablet", "Smartwatch", "Speakers"]

# Fixed set of transactions for deterministic results

transactions = [

["Laptop", "Phone", "Charger"],

["Phone", "Headphones", "Charger", "Smartwatch"],

["Laptop", "Tablet", "Speakers"],

["Monitor", "Keyboard", "Mouse"],

["Laptop", "Monitor", "Keyboard", "Mouse"],

["Phone", "Smartwatch"],

["Headphones", "Charger"],

["Laptop", "Headphones", "Charger"],

["Monitor", "Tablet", "Speakers"],

["Laptop", "Phone", "Headphones", "Smartwatch"],

["Laptop", "Phone", "Mouse", "Keyboard"],

] \* 91 # Repeat the fixed set 91 times to have 1001 transactions

```

1. Use the Apriori algorithm to find all frequent itemsets with a minimum support threshold of 0.185.

2. From the frequent itemsets, generate association rules with a minimum confidence threshold of 0.6. List the top 5 rules with the highest lift.

## Problem Guidance

### Part A: Theoretical Aspects

#### 1. Definitions:

Approach:

- Use your notes, textbooks, or any learning materials to recall the definitions.

- Write the definitions in your own words, ensuring clarity and comprehensiveness.

- Make sure to mention the formula for each term.

Expected Answer:

- A brief description for each term.

- Mathematical representation/formula for each term.

- Approximate length: 2-4 sentences per term.

#### 2. Apriori Algorithm:

Approach:

- Recall the steps of the Apriori algorithm.

- Think about the significance of the pruning step and its impact on efficiency.

- Use bullet points or numbering to describe the steps in a sequential manner.

Expected Answer:

- A step-by-step breakdown of the algorithm.

- Explanation of the importance of pruning in the algorithm.

- Approximate length: 7-10 sentences.

#### 3. Calculations:

Approach:

- Use the formulas for support and confidence to compute the required values.

- Break down the task into smaller steps:

1. Count the number of transactions containing the given items.

2. Divide by the total number of transactions to get the support.

3. Use the support values to compute the confidence.

- Clearly show each step of your calculation for clarity.

Expected Answer:

- Detailed calculations showing each step.

- Final values for support and confidence.

- Approximate length: 5-7 sentences.

---

### Part B: Implementation

#### 1. Generate the transactions:

Approach:

- Use the provided code snippet to generate the transactions.

- Write your own small script to find the transactions that exceed the minimum support threshold.

- Make sure you understand what each line of code does.

#### 2. Use the Apriori algorithm and Generate the top 5 association rules:

Approach:

- Begin by transforming the transactions into a format suitable for the Apriori algorithm.

- Use the `mlxtend` library's functions to compute frequent itemsets and association rules.

- Follow the provided code and ensure you understand each step.

- Sort the association rules based on the lift value.

- Extract the top 5 rules.

- Ensure your answer is clear and readable.

Expected Answer:

- A DataFrame containing the frequent itemsets with their support values.

- Approximate length: Depends on the generated transactions, but expect around 10-20 itemsets.

- A list or table of the top 5 association rules with the highest lift.

- Each rule should show the antecedents, consequents, support, confidence, and lift.

- Approximate length: 5 rules with details for each rule.

---

General Tips:

- Always ensure you understand the problem statement and what is being asked.

- Break down the problem into smaller tasks and tackle each one methodically.

- For the theoretical part, ensure clarity and precision in your answers.

- For the implementation part, test each section of your code to ensure it works as expected.

- Finally, review your answers to ensure they align with the problem's requirements.

# Question 2

#### Problem Statement

You are a data scientist working for a healthcare research institute. You've been provided a dataset containing medical records of patients, including their age, weight, height, and blood pressure. The institute is interested in running some advanced machine learning algorithms on this dataset, but before doing so, they want to ensure the data is clean and normalized.

## Part A: Data Preprocessing

1. Identify and handle any missing values in the dataset.

2. Detect and correct any outliers in the blood pressure column using the IQR method.

3. Encode any categorical variables present in the dataset.

## Part B: Data Normalization

1. Normalize the age, weight, and height columns using Min-Max normalization.

2. Normalize the blood pressure column using Z-score normalization.

3. Explain the difference between the two normalization methods and why you might choose one over the other in different scenarios.

Given dataset (sample):

A table with numbers and letters

Description automatically generated

Note: "-" indicates missing values.

---

## Problem Guidance:

### Part A: Data Preprocessing

#### 1. Handling Missing Values:

- Begin by identifying columns with missing values.

- Decide on an approach to handle them (e.g., removing rows, filling with mean/median, or using imputation techniques).

- Implement the chosen approach.

Expected Answer:

- A statement identifying which columns had missing values and how you decided to handle them.

- The dataset after handling missing values.

- Approximate length: 3-5 sentences and the modified dataset.

#### 2. Outlier Detection and Correction:

- Use the IQR method to detect outliers in the blood pressure column.

- Decide on a method to handle outliers (e.g., removing them or capping them at certain values).

- Implement the chosen method.

Expected Answer:

- Statement identifying detected outliers and the chosen method to handle them.

- The dataset after handling outliers.

- Approximate length: 3-5 sentences and the modified dataset.

#### 3. Encoding Categorical Variables:

- Identify any categorical columns in the dataset.

- Choose an encoding technique (e.g., one-hot encoding, label encoding).

- Implement the chosen encoding method.

Expected Answer:

- A statement identifying which columns were encoded and the method used.

- The dataset after encoding.

- Approximate length: 2-4 sentences and the modified dataset.

### Part B: Data Normalization

#### 1. Min-Max Normalization:

- Apply Min-Max normalization to the age, weight, and height columns.

- The formula is: A close up of a word

Description automatically generated

Expected Answer:

- The dataset with age, weight, and height columns normalized.

- Approximate length: The modified dataset.

#### 2. Z-score Normalization:

- Apply Z-score normalization to the blood pressure column.

- The formula is: A close-up of words

Description automatically generated

Expected Answer:

- The dataset with the blood pressure column normalized.

- Approximate length: The modified dataset.

#### 3. Difference Between Normalization Methods:

- Explain the concept behind each normalization method.

- Discuss scenarios where one might be preferred over the other.

Expected Answer:

- A clear explanation of Min-Max and Z-score normalization.

- Discussion on the pros and cons of each method and their ideal use cases.

- Approximate length: 5-7 sentences.

---

General Tips:

- Start by thoroughly reading and understanding the problem statement.

- Before jumping into solutions, plan your approach by breaking down tasks.

- For theoretical parts, ensure clarity in your explanations. Use your own words to demonstrate understanding.

- For the implementation parts, test your solutions to ensure correctness.

- Regularly review the problem statement to ensure you're on track and meeting the requirements.

# Question 3

#### Problem Statement

You are a data scientist working for a wine distributor. The company sources wines from three different vineyards in Italy. They've collected data on the chemical composition of wines from each vineyard, including attributes like alcohol content, malic acid, ash, and more. Each wine belongs to one of the three vineyards, and the company wants to develop a machine learning model to predict the origin of the wine based on its chemical composition.

The data set is available here:

[Wine - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/109/wine)

## Part A: Cross Validation

1. Explain the concept of k-fold cross-validation and its importance in model evaluation.

2. Given the wine dataset with 178 records, implement 5-fold cross-validation. Describe the size of your training and test sets in each iteration.

3. Discuss the advantages and disadvantages of using a larger k in k-fold cross-validation.

## Part B: Classifier Evaluation

1. After training your model on the training set, suppose you get the following confusion matrix on one of the test sets for the three classes (Vineyard 1, Vineyard 2, Vineyard 3):

A white background with black text

Description automatically generated

Calculate and interpret the following metrics for Vineyard 1 based on the confusion matrix:

- Accuracy

- Precision

- Recall (Sensitivity)

- F1-Score

2. Discuss the importance of evaluating a classifier using multiple metrics rather than solely relying on accuracy, especially in the context of wine classification.

3. The distributor is particularly concerned about misclassifying a wine from Vineyard 1 as one from Vineyard 2 or Vineyard 3. Which metric should be the primary focus, and why?

## Problem Guidance

### Part A: Cross Validation

#### 1. Understanding k-fold Cross Validation:

Approach:

- Begin by defining what cross-validation is in the context of machine learning and model evaluation.

- Elaborate on the specific process of k-fold cross-validation: how data is partitioned, how training and testing are done iteratively, and why this approach is beneficial.

Expected Answer:

- A comprehensive explanation of the k-fold cross-validation technique.

- Justification for its importance in model evaluation.

- Approximate length: 5-7 sentences.

#### 2. Implementing 5-fold Cross Validation:

Approach:

- Explain how you'd divide the wine dataset into 5 equally sized or nearly equally sized partitions.

- Describe how in each iteration, one partition will serve as the test set while the others will serve as the training set.

Expected Answer:

- A step-by-step breakdown of the 5-fold cross-validation process on the wine dataset.

- Specific sizes for training and test sets for each of the 5 iterations.

- Approximate length: 5-7 sentences.

#### 3. Discussing the Choice of k:

Approach:

- Reflect on the implications of choosing different values of k in k-fold cross-validation.

- Consider the trade-offs in terms of data utilization, computational cost, and evaluation variance.

Expected Answer:

- A discussion of the pros and cons of increasing k in k-fold cross-validation.

- Approximate length: 4-6 sentences.

---

### Part B: Classifier Evaluation

#### 1. Calculating Metrics for Vineyard 1:

Approach:

- Start by defining each metric's formula: Accuracy, Precision, Recall, and F1-Score.

- Use the provided confusion matrix to extract values: True Positives, False Positives, True Negatives, and False Negatives for Vineyard 1.

- Plug in these values into each metric's formula to compute the required values.

Expected Answer:

- Calculations and results for each metric: Accuracy, Precision, Recall, and F1-Score.

- Brief interpretations or implications of each metric's value in the context of wine classification.

- Approximate length: 8-10 sentences.

#### 2. Importance of Multiple Metrics:

Approach:

- Discuss the limitations of relying solely on accuracy as a performance measure.

- Elaborate on scenarios or situations in the wine dataset where accuracy might be misleading.

- Highlight the significance of other metrics like Precision and Recall in providing a more comprehensive view of the classifier's performance.

Expected Answer:

- A discussion on the potential pitfalls of only considering accuracy.

- Emphasis on the importance of looking at multiple evaluation metrics for a holistic view of classifier performance.

- Approximate length: 4-6 sentences.

#### 3. Focus on False Negatives:

Approach:

- Understand the implications of False Negatives in the context of the problem (misclassifying a wine from Vineyard 1).

- Identify which metric directly evaluates the rate of False Negatives and explain why it's crucial.

Expected Answer:

- Identification of the most relevant metric (Recall) given the distributor's concern.

- Explanation of why this metric should be prioritized in this context.

- Approximate length: 2-4 sentences.

---

General Tips:

- Always start by thoroughly reading and understanding the problem statement.

- For theoretical parts, clarity in explanations is key. Use your own words to demonstrate a deep understanding.

- For calculation or implementation parts, ensure each step is logical and well-justified.

- Regularly refer back to the problem statement to ensure all aspects are addressed.

- After solving, review your answers to ensure they are comprehensive and address the problem's requirements.

# Question 4

#### Problem Statement

You are a data scientist at a major e-commerce company. The company is interested in offering personalized discounts to its customers to improve sales. They believe that a customer's likelihood to purchase an item can be predicted based on certain features like the time of day they browse, their browsing device, and the category of items they usually explore.

You've decided to use a decision tree classifier to model this problem. The dataset provided to you contains the following features:

1. Time of Day: Morning, Afternoon, Evening, Night

2. Device: Mobile, Desktop, Tablet

3. Category: Electronics, Clothing, Groceries, Books

4. Purchase: Yes, No

Using a sample of this dataset:

A screenshot of a phone

Description automatically generated

## Part A: Decision Trees

1. Based on the provided data, which feature would you choose as the root node of your decision tree? Explain your choice.

2. Construct a simple decision tree up to a depth of 1 using the feature you've selected. What can you infer about the buying behavior of customers from this tree?

## Part B: Information Gain

1. Calculate the entropy of the target variable "Purchase".

2. Compute the information gain for each feature to justify your choice for the root node in Part A.

3. Explain the importance of information gain in the process of constructing decision trees.

---

## Problem Guidance:

### Part A: Decision Trees

#### 1. Choosing the Root Node:

Approach:

- Review the dataset and observe the patterns in which features might determine the "Purchase" decision.

- Consider features that provide a clear distinction between the "Yes" and "No" outcomes.

Expected Answer:

- Identification of one feature as the potential root node.

- Justification based on the observed patterns in the dataset.

- Approximate length: 3-5 sentences.

#### 2. Constructing the Decision Tree:

Approach:

- Using the chosen root node, segment the data based on its distinct values.

- For each segment, determine the most frequent "Purchase" outcome.

Expected Answer:

- A simple decision tree diagram or description up to a depth of 1.

- Interpretations or inferences about customer buying behavior based on the tree.

- Approximate length: 4-6 sentences.

---

### Part B: Information Gain

#### 1. Calculating Entropy:

Approach:

- Understand the formula for entropy: A black emoticon with a line

Description automatically generated where p+ and p- are the probabilities of the positive and negative classes, respectively.

- Calculate the entropy of the "Purchase" column based on the provided data.

Expected Answer:

- Computation of the entropy of the target variable.

- Approximate length: 2-3 sentences.

#### 2. Computing Information Gain:

Approach:

- Understand the formula for information gain: A black and white logo

Description automatically generated where S is the original set and Sv is the subset of S for which attribute A has value v.

- Compute the information gain for each feature relative to the target variable "Purchase".

Expected Answer:

- Computation of the information gain for each feature.

- Justification of the chosen root node based on the highest information gain.

- Approximate length: 5-7 sentences.

#### 3. Importance of Information Gain:

Approach:

- Reflect on how information gain measures the effectiveness of an attribute in classifying data.

- Discuss its role in constructing decision trees and ensuring that the most significant attributes are considered first.

Expected Answer:

- Explanation of the concept of information gain.

- Its significance in the decision tree construction process.

- Approximate length: 4-6 sentences.

---

General Tips:

- Start with a clear understanding of the problem domain (e-commerce) and the dataset.

- For theoretical parts, ensure clarity in explanations.

- For calculation parts, always explain your steps and approach before diving into computations.

- After solving, review your answers for completeness and coherence.