CLASS TEST 1 SET 2 ANSWER FILE

PAPER CODE-MCA305B

PAPER NAME-BIG DATA ANALYSIS

1.

i) **Define Supervised ML:** Answer: b) Trained with labelled data

ii) **Define Confusion Matrix:** Answer: b) Measures ML Model's performance

iii) **The primary aspect of Apriori Algorithm is to provide:** Answer: a) Frequent Itemset

iv) **K-means is a:** Answer: d) Unsupervised ML Model

v) **State about the master job manager to oversee:** Answer: d) All of these

vi) **Demonstrate about categories of clustering methods:** Answer: d) All of these (Partitioning methods, Hierarchical methods, Grid-based methods)

vii) **Describe about semi-structured data, represented in an:** Answer: a) XML file

viii) **Justify which of the following function is used for k-means clustering:** Answer: a) k-means

Short Answer Type Questions:

2. **Explain support count and lift:**
    - Support Count: Support count is a measure used in association rule mining. It calculates the number of times a particular itemset or rule appears in the dataset. It helps identify the frequency of occurrence of a rule or itemset. Higher support count values indicate a stronger association between items.
    - Lift: Lift is a measure used to evaluate the strength of association between items in an association rule. It compares the likelihood of items being purchased together in relation to their individual purchase probabilities. A lift value greater than 1 indicates a positive association, meaning that the items are more likely to be purchased together than individually. A lift value of 1 implies no association.
3. **The Twitter data satisfies characteristics of big data:** Twitter data often satisfies the characteristics of big data because it is generated at a high velocity (real-time tweets), in large volumes (millions of tweets per day), and comes in diverse formats (text, images, videos). Additionally, Twitter data can exhibit veracity challenges, as it may contain noise, spam, and unreliable information. Analyzing Twitter data typically involves big data technologies and analytics due to its volume and real-time nature.
4. **Differentiate between Recall and Precision:**
    - Recall (Sensitivity): Recall is a measure of the model's ability to correctly identify all relevant instances in a dataset. It is calculated as the ratio of true positives to the sum of true positives and false negatives. High recall means the model is good at capturing all relevant instances, but it may have more false positives.
    - Precision: Precision is a measure of the model's ability to correctly identify only relevant instances out of all instances it identifies as positive. It is calculated as the ratio of true positives to the sum of true positives and false positives. High precision

indicates that when the model predicts a positive instance, it is usually correct, but it may miss some relevant instances.

5. **Illustrate the features of a Confusion Matrix:** A Confusion Matrix is a table used to evaluate the performance of a classification model. It includes the following features:
    - True Positives (TP): Correctly predicted positive instances.
    - True Negatives (TN): Correctly predicted negative instances.
    - False Positives (FP): Incorrectly predicted positive instances.
    - False Negatives (FN): Incorrectly predicted negative instances.
6. **Pros and cons of K-means clustering algorithm:**
    - Pros:
        - Simple and easy to implement.
        - Efficient for large datasets.
        - Suitable for a wide range of applications.
        - Provides hard clustering, where each data point belongs to one cluster.
    - Cons:
        - Sensitive to the initial choice of centroids.
        - Assumes clusters are spherical and equally sized, which may not be the case in some datasets.
        - May converge to a local minimum.
        - Not suitable for non-linearly separable data.
7. **Illustrate KNN ML method with an example:** K-Nearest Neighbors (KNN) is a supervised machine learning method used for classification and regression. In KNN, a data point is classified based on the majority class among its k-nearest neighbors. For example, in a KNN classification problem, given a dataset of labeled flowers with features like petal length and width, KNN can be used to classify an unlabeled flower by looking at the class of the k-nearest labeled flowers in feature space.