MACHINE LEARNING – WORKSHEET
(CLUSTERING)

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. Movie Recommendation systems are an example of:
   1. Classification
   2. Clustering
   3. Reinforcement Learning
   4. Regression
      Options:
      a. 2 Only
      b. 1 and 2
      c. 1 and 3
      d. 2 and 3 ✓
      e. 1, 2 and 3
      f. 1, 2, 3 and 4

2. Sentiment Analysis is an example of:
   1. Regression
   2. Classification
   3. Clustering
   4. Reinforcement Learning
      Options:
      a. 1 Only
      b. 1 and 2
      c. 1 and 3
      d. 1, 2 and 3
      e. 1, 2 and 4 ✓
      f. 1, 2, 3 and 4

3. Can decision trees be used for performing clustering?
      a. True ✓
      b. False

4. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:
      a. Capping and flooring of variables
      b. Removal of outliers
         Options:
         a. 1 only ✓
         b. 2 only
         c. 1 and 2
         d. None of the above

5. What is the minimum no. of variables/ features required to perform clustering?
    a. 0
    b. 1 ✓
    c. 2
    d. 3

6. For two runs of K-Mean clustering is it expected to get same clustering results?
    a. Yes
    b. No ✓

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means
    a. Yes ✓
    b. No
    c. Can't say
    d. None of these

8. Which of the following can act as possible termination conditions in K-Means?
    1. For a fixed number of iterations.
    2. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
    3. Centroids do not change between successive iterations.
    4. Terminate when RSS falls below a threshold.
        Options:
            a. 1, 3 and 4
            b. 1, 2 and 3
            c. . 1, 2 and 4
            d. . All of the above ✓

9. Which of the following can act as possible termination conditions in K-Means?
    1. K- Means clustering algorithm
    2. Agglomerative clustering algorithm
    3. Expectation-Maximization clustering algorithm
    4. Diverse clustering algorithm
        Options:
            a. 1 only
            b. 2 and 3
            c. 2 and 4
            d. 1 and 3 ✓
            e. 1,2 and 4
            f. All of the above

10. Which of the following algorithms is most sensitive to outliers?
    a. K-means clustering algorithm ✓
    b. K-medians clustering algorithm
    c. K-modes clustering algorithm
    d. K-medoids clustering algorithm

11. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):
    1. Creating different models for different cluster groups.
    2. Creating an input feature for cluster ids as an ordinal variable.
    3. Creating an input feature for cluster centroids as a continuous variable.
    4. Creating an input feature for cluster size as a continuous variable.
       Options:
       a. 1 only
       b. 1 and 2
       c. 1 and 4
       d. 3 only
       e. 2 and 4 ✓
       f. All of the above

12. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?
       a. Proximity function used
       b. of data points used
       c. of variables used
       d. B and c only
       e. All of the above ✓

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly

13. Is K sensitive to outliers?
14. Why is K means better?
15. Is K means a deterministic algorithm?

13 answer: K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. K-medoids clustering is a variant of K-means that is more robust to noises and outliers. The group of points in the right form a cluster, while the rightmost point is an outlier.

14 answer: Other clustering algorithms with better features tend to be more expensive. In this case, k-means becomes a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where other clustering algorithms can be applied. K-means is the simplest. To implement and to run.

15 answer: k-means clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results. However, to ensure consistent results, FCS Express performs k-means clustering using a deterministic method.