**Q1 to Q11 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. Regression Options:
      1. 2 Only
      2. 1 and 2
      3. 1 and 3
      4. 2 and 3
2. Sentiment Analysis is an example of:
   1. Regression
   2. Classification
   3. Clustering
   4. Reinforcement Options:
      1. 1 Only
      2. 1 and 2
      3. 1 and 3
      4. 1, 2 and 4
3. Can decision trees be used for performing clustering?
4. True
5. False
6. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:
   1. Capping and flooring of variables
   2. Removal of outliers Options:
      1. 1 only
      2. 2 only
      3. 1 and 2
      4. None of the above
7. What is the minimum no. of variables/ features required to perform clustering?
8. 0
9. 1
10. 2
11. 3
12. For two runs of K-Mean clustering is it expected to get same clustering results?
13. Yes
14. No
15. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?
16. Yes
17. No
18. Can't say
19. None of these
20. 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 witha bad local minimum.
    3. Centroids do not change between successive iterations.
    4. Terminate when RSS falls below a threshold. Options:
       1. 1, 3 and 4
       2. 1, 2 and 3
       3. 1, 2 and 4
       4. All of the above
21. Which of the following algorithms is most sensitive to outliers?
22. K-means clustering algorithm
23. K-medians clustering algorithm
24. K-modes clustering algorithm
25. K-medoids clustering algorithm
26. 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:
       1. 1 only
       2. 2 only
       3. 3 and 4
       4. All of the above
27. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?
28. Proximity function used
29. of data points used
30. of variables used
31. All of the above

Q12 to Q14 are subjective answers type questions, Answers them in their own words briefly

1. Is K sensitive to outliers?
2. Why is K means better?
3. Is K means a deterministic algorithm?

12. K-means is highly sensitive to outliers. Outliers can influence the final decision and are removed to produce quality output. As the mean, as a statistic output is generally sensitive to outliers, k-means a method is susceptible to outliers.

For example; The means of 2,2,2,2,2,2,4,4,5,5 is 3

If we add a 50 to the means, the output will be 8.

The output is 8, which is larger than *any* of the other values.

Since in k-means, means is used a lot, one ends up with a lot of outlier-sensitive calculations.

13. K-means is simple to implement and is scalable to large data sets.

It guarantees convergence and is a good start for selecting the positions of centroids.

It adapts easily new examples and is good at generalizing to clusters of different sizes and shapes.

14. K-Means is **non-deterministic** in nature which means that running the algorithm several times on the same data set, can give different results. In this case, the compiler cannot solve the problem in polynomial time and does not know the next step as some problems have a great degree of randomness than others.