MACHINE LEARNING WORKSHEET – 2

1. Movie Recommendation systems are an example of -
i) Classification
ii) Clustering
iii) Regression
Answer - **(d) 2 and 3**

2. Sentiment Analysis is an example of -
i) Regression
ii) Classification
iii) Clustering
iv) Reinforcement
Answer - **(d) 1, 2 and 4**

3. Can decision trees be used for performing clustering?
Ans **(a) True**

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 -
i) Capping and flooring of variables
ii) Removal of outliers
Answer - **(a) 1 only**

5. What is the minimum no. of variables/ features required to perform clustering?
Answer - **(b) 1**

6. For two runs of K-Mean clustering is it expected to get same clustering results?
Answer - **(b) No**

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?
Answer - **(a) Yes**

8. Which of the following can act as possible termination conditions in K-Means?
i) For a fixed number of iterations
ii) Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum
iii) Centroids do not change between successive iterations
iv) Terminate when RSS falls below a threshold
Answer - **(d) All of the above**

9. Which of the following algorithms is most sensitive to outliers?
Answer - **(a) K-means clustering algorithm**

10. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning) -
i) Creating different models for different cluster groups
ii) Creating an input feature for cluster ids as an ordinal variable
iii) Creating an input feature for cluster centroids as a continuous variable
iv) Creating an input feature for cluster size as a continuous variable
Answer - **(d) All of the above**

11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?
Answer - **(d) All of the above**

12. Is K sensitive to outliers?

Answer - **Centroids are calculated via the K-means clustering algorithm, which then iterates until the best centroid is discovered. It is assumed that the number of clusters is known. The flat clustering algorithm is yet another name for it. The letter "K" in K-means stands for the number of clusters that the algorithm identified from the data.**

According to this strategy, data points are grouped into clusters so that the total of their squared distances from the centroid is as little as it can be. Important to notice is that less unique data points within clusters result in more identical data points within the same cluster.

K-means centroids are iteratively calculated by averaging the cluster's data points. Outliers have an impact on averages. As a result, before training our model, it is vital to identify and eliminate outliers.

13. Why is K means better?

Answer - **The k-means clustering algorithm benefits are as follows -**

- **Fast and reliable**
- **Simple to comprehend**
- **Comparably effective**
- **The best outcomes are obtained when data sets are distinct.**
- **Create more compact clusters.**
- **The cluster varies as centroids are recomputed.**
- **Flexible**
- **Better computational cost, Simple to interpret**
- **Enhances Accuracy**
- **uses spherical clusters better**

14. Is K means a deterministic algorithm?

Answer - **Randomness-based clustering algorithms frequently produce inconsistent results for the same dataset when run many times. The non-deterministic character of algorithms, like the K-Means clustering algorithm, restricts their use in fields like predicting cancer subtypes from gene expression data.**

Comparing the outcomes of such algorithms against those of other algorithms in a reasonable manner is challenging. The random selection of data points used to create the initial centroids of K-Means accounts for its non-deterministic nature.