Machine Learning (Assignment 3)

1. d) All of the above

2. d) None

3. c) Reinforcement learning and unsupervised learning

4. b) The tree representing how close the data points are to each other

5. d) None

6. c) K-nearest neighbour is same as K-means

7. d) 1, 2 and 3

8. a) 1 only

9. a) 2

10. b) Given a database of information about your users, automatically group them into different market segments.

11. a)

12. b)

13. Clustering can be used in various applications:
    a) **Outlier detection:** We can use it to find outliers. We can easily detect unusual behaviour of the customers so that we can use it to detect fraud.
    b) **Customer segmentation:** We can identify customers who has similar activities and cluster them based on those similar instances they have. So, we can use this data to focus more on what each group needs and provide their interest and improve services. We can recommend them certain items based on their group.
    c) **Data analysis:** If we cluster data, it is way easier to analyse since we would already know their similarities.
    d) **For search engines:** If a person search for some item online, the search engine will provide all the related items from the cluster it has already formed.

14. To improve clustering performance:
    - We need to specify the number of clusters beforehand.
    - Run the algorithm multiple times to avoid a sub-optimal solution
    - K-means does not behave very well when the clusters have varying sizes, different densities, or non-spherical shapes. We can use k-means++ that tends to select centroids that are distant from one another and this makes the k-means algorithm much less likely to converge to a sub-optimal solution.
    - Instead of using the full dataset at each iteration, the algorithm is capable of using **mini-batches**, moving the centroids just slightly at each iteration. This speeds up the algorithm typically by a factor of 3 or 4 and makes it possible to cluster huge datasets that do not fit in memory.