LAB Assignment 6
CSL 2050

By :- Akshat Jain B21CS005

1. We preprocess the data and remove column 0,7,10 from the X part of the dataset as column 0 just indexes, column 10 is the target variable, and columns 1 and 7 are highly correlated
    1.1. We perform the above stated tasks and visualise the k-means clustering algorithm for all pairwise columns
    1.2. We calculate the silhoutte score and on the basis of that find that the best value for k is generally 2 or 3.
    1.3. We then use the elbow method and determine that k = 3 is the best value.
    1.4. We go for bagging as bagging reduces variance by taking average prediction. Bias may remain unchanged or even increase a little bit due to increase in randomness.
2. N/A
3. We do the Above stated tasks
    3.1. We preprocess the data and apply standard scaling on all the features except region and channel as they are categorical features
    3.2. We find that maximum covariance is between the features ‘Groceries’ and ‘Detergents_paper’
    3.3. We apply DBSCAN and visualise it for a range of eps(distance) values.
    3.4. We apply KNN on the same 2 features and target as channel. We see that in DBSCAN due to the nature of this data it was very difficult to differentiate between clusters and there was a huge amount of data points classified as noisy. Whereas in the case of KNN we see that the accuracy here is a little low due to the fact that many of the points lie on the wrong side of their respective decision boundaries since the feature values of

these datapoints are intertwined with each other to a large extent