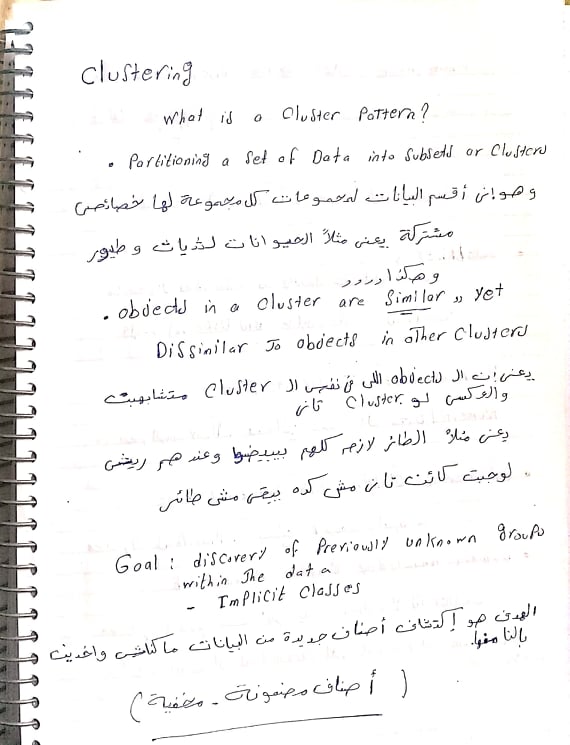
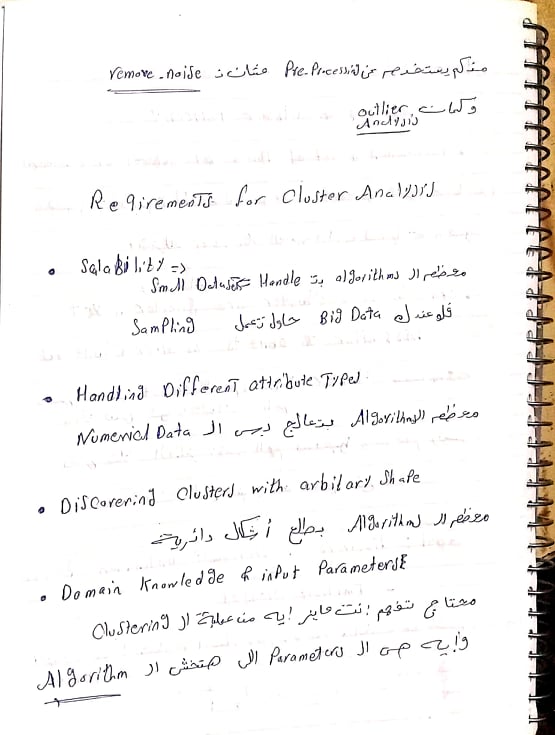
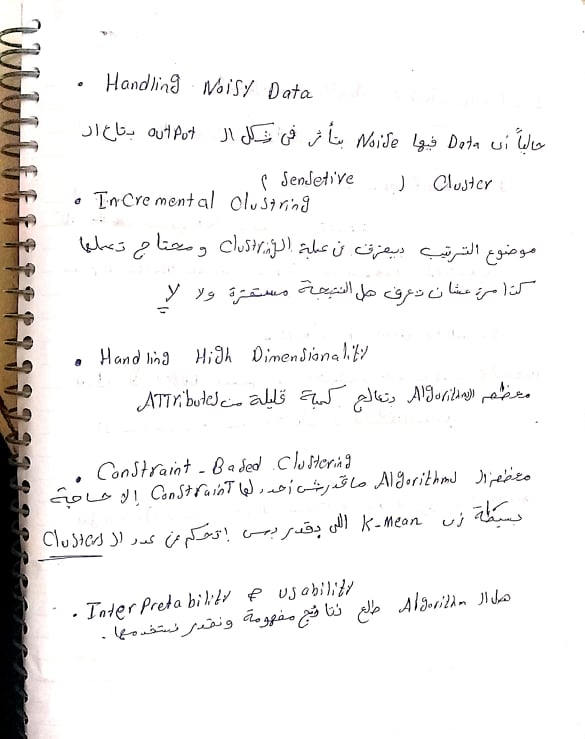


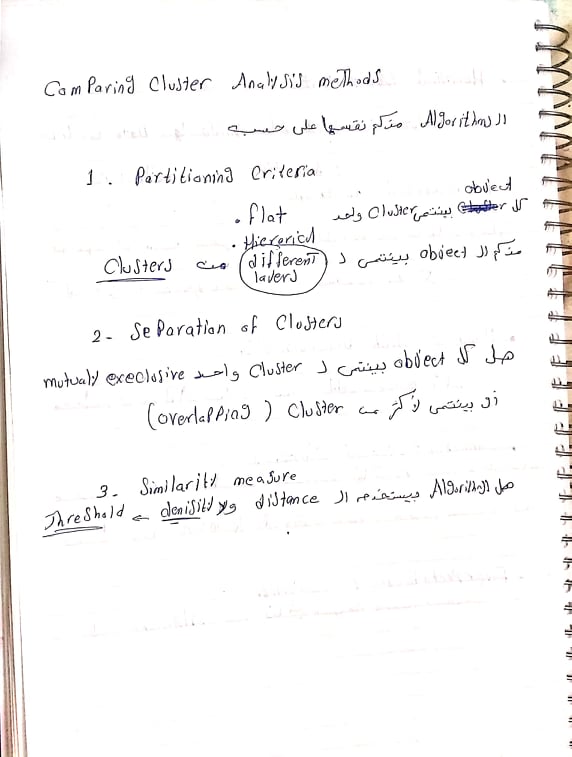
Ahmed Khalil

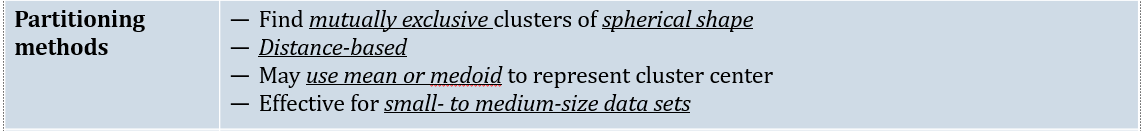
[Company name]  [Company address]

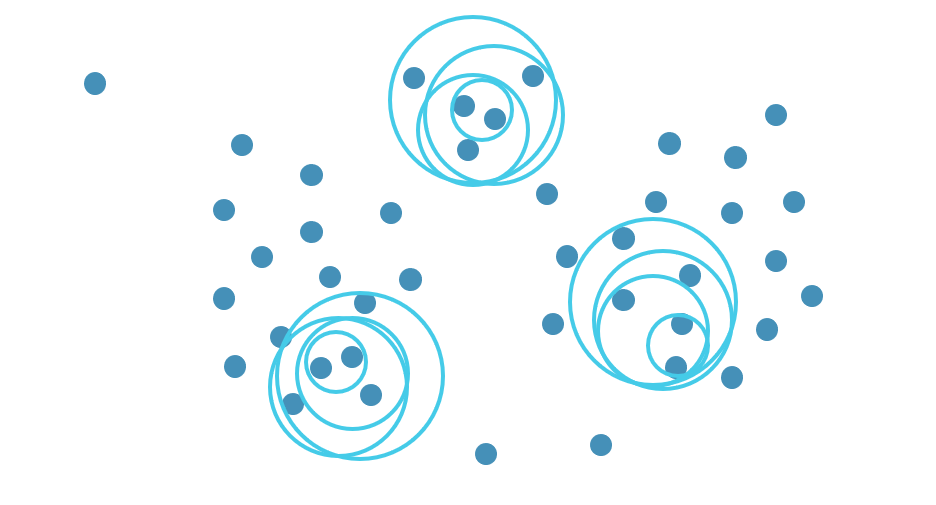
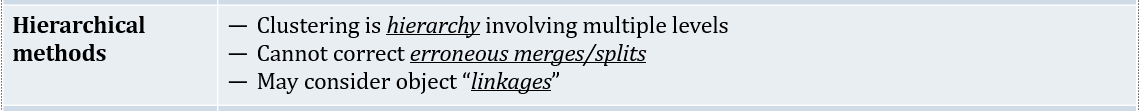
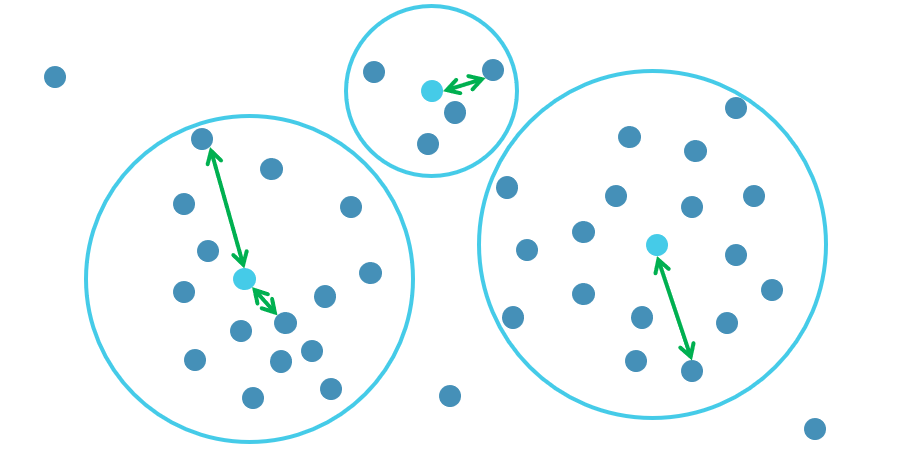
Clustering Pattern

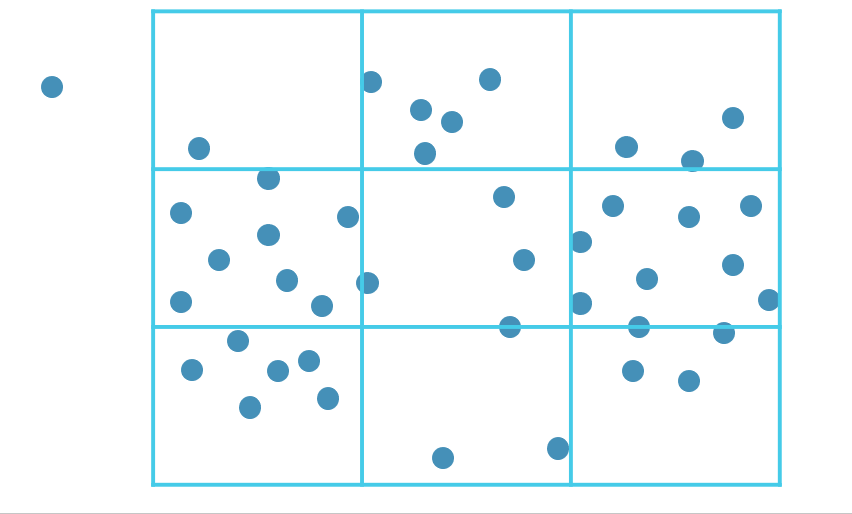
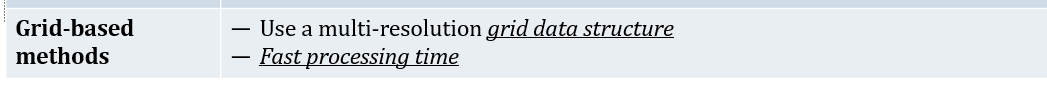
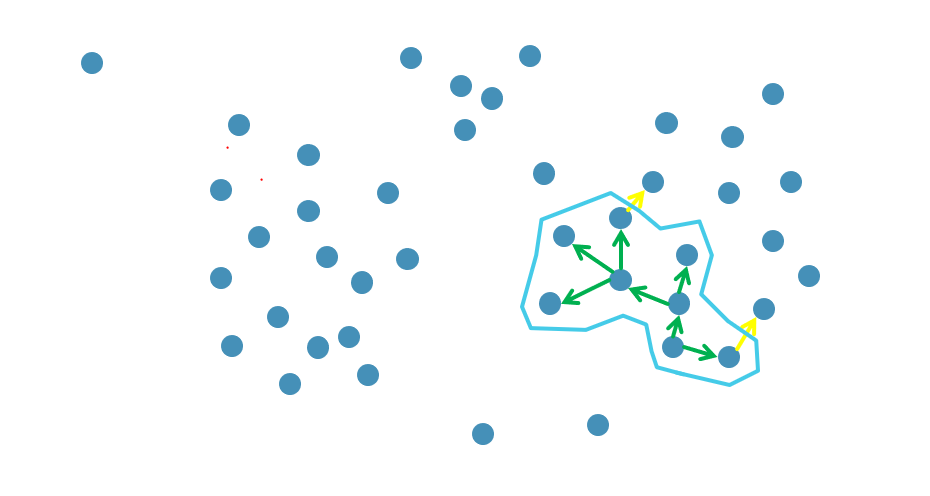
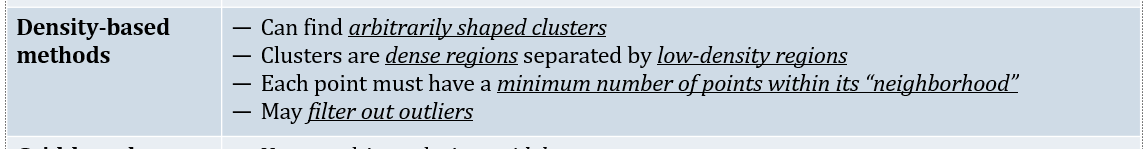




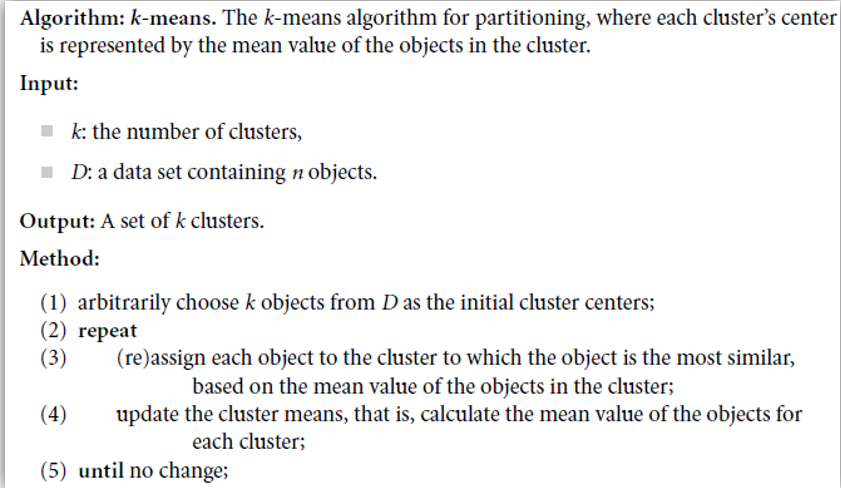
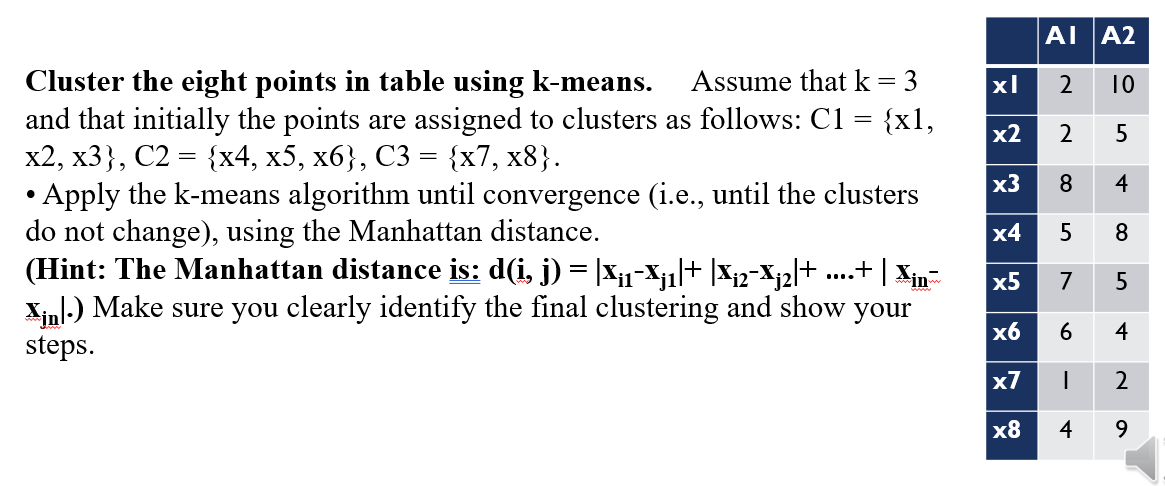


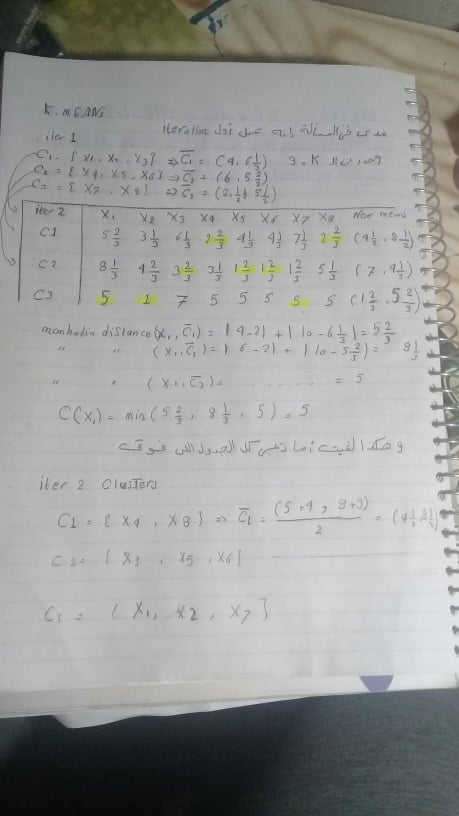




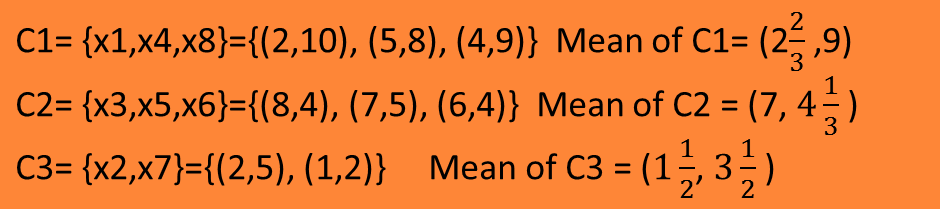


# Partitioning Methods {K-Means , K-Medoid}



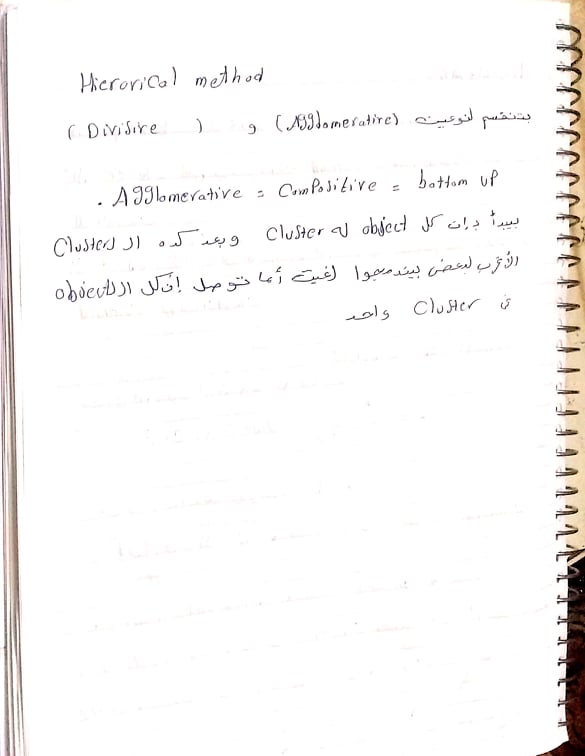


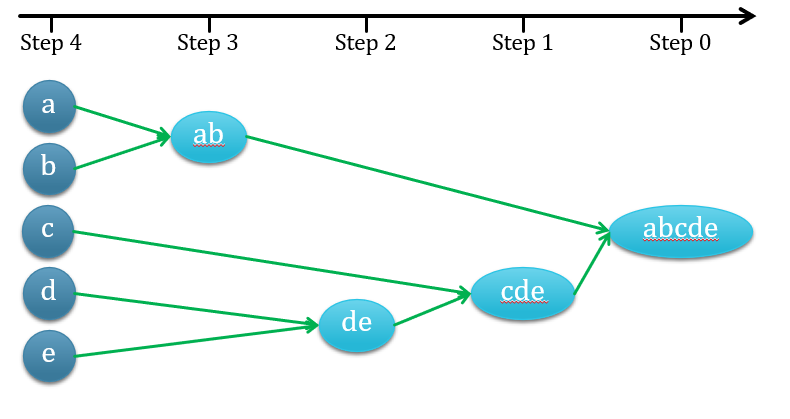
* و هتفضل شغال كده في iterations لغيت اما تلاقى ال clusters شكلها ثبتت و ده الحل

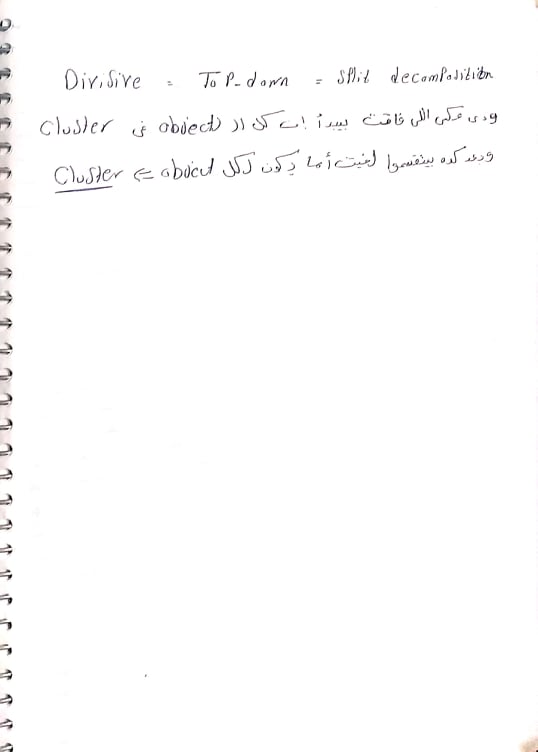


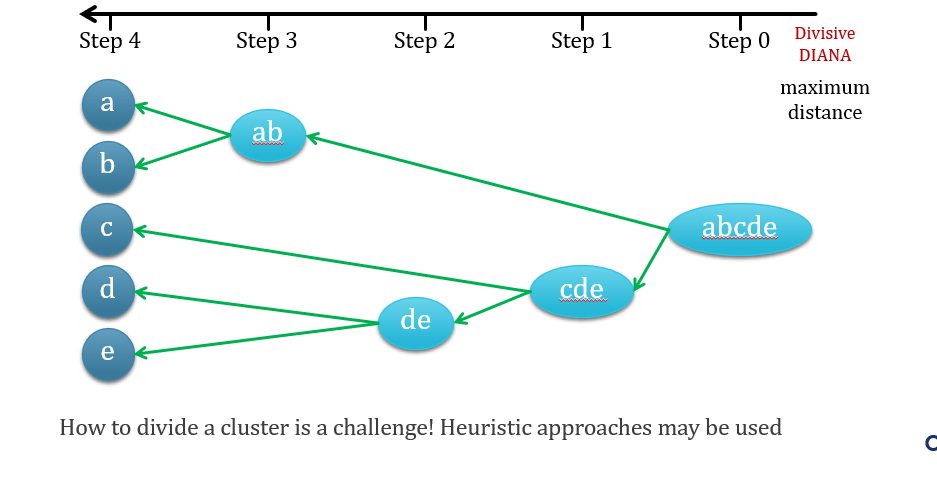
# Hierarchical Methods {Agglomerative vs Devisive}

وبنطلع فيها ال clusters على هيئة شجرة متتابعة و فيه منها نوعين من ال methods

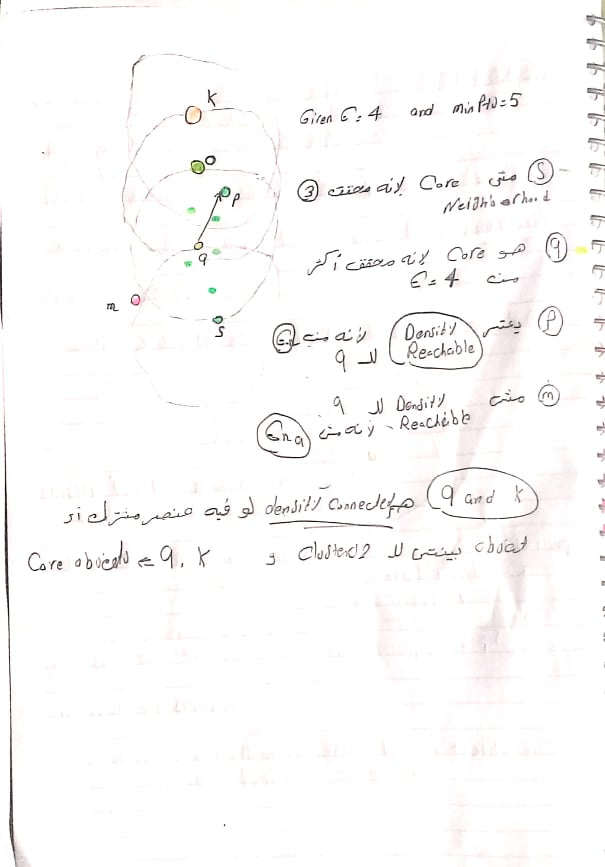


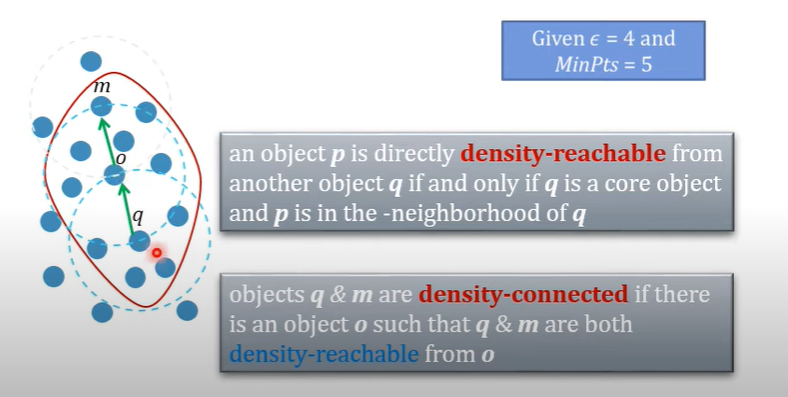


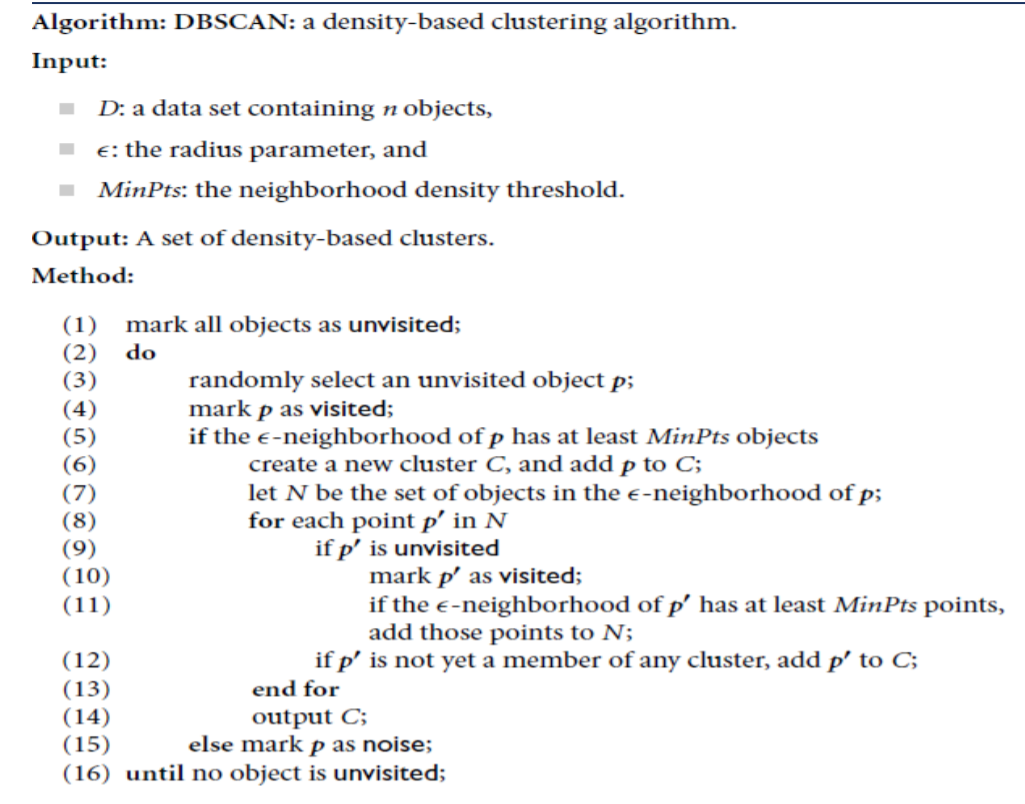




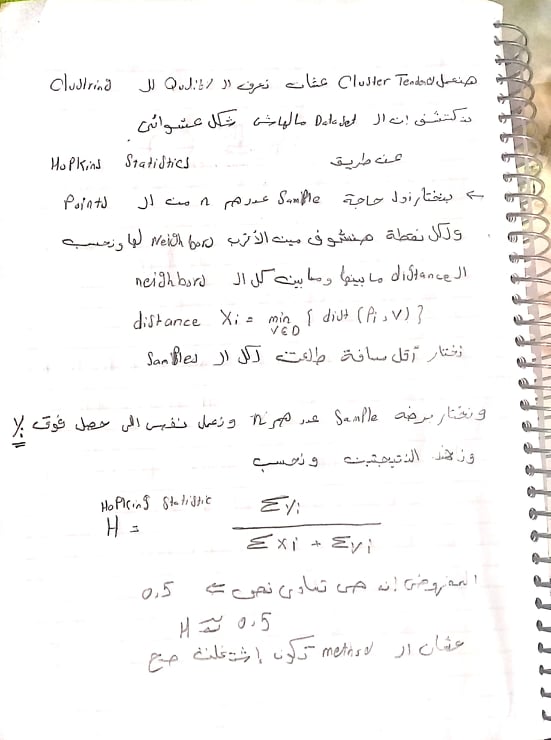
# Density Based Methods {DBSCAN}

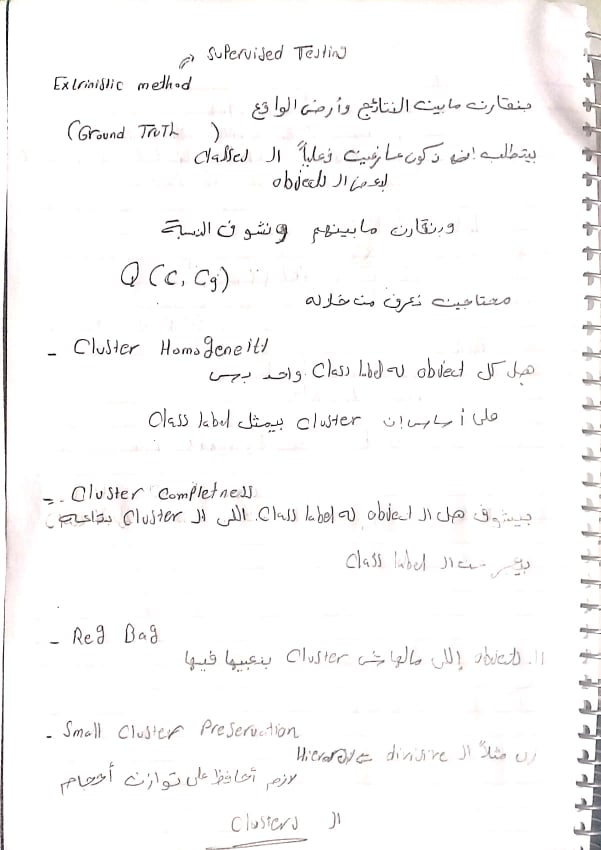


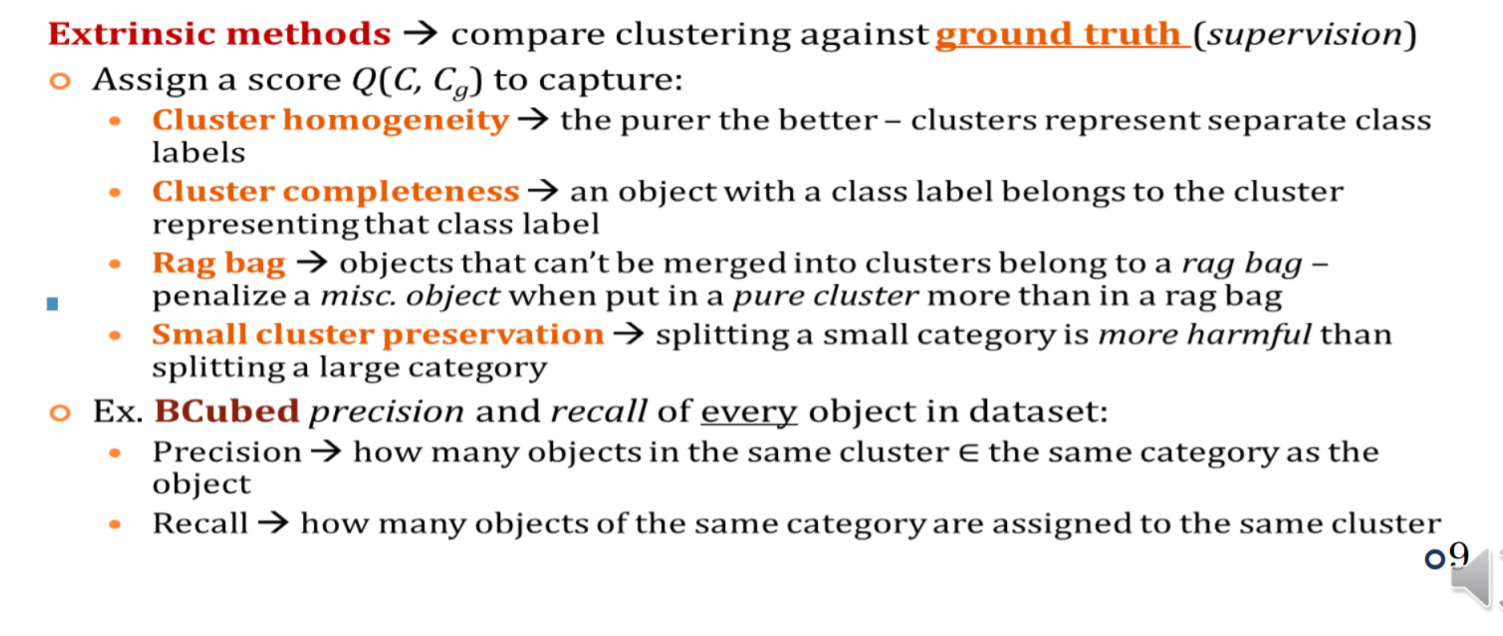
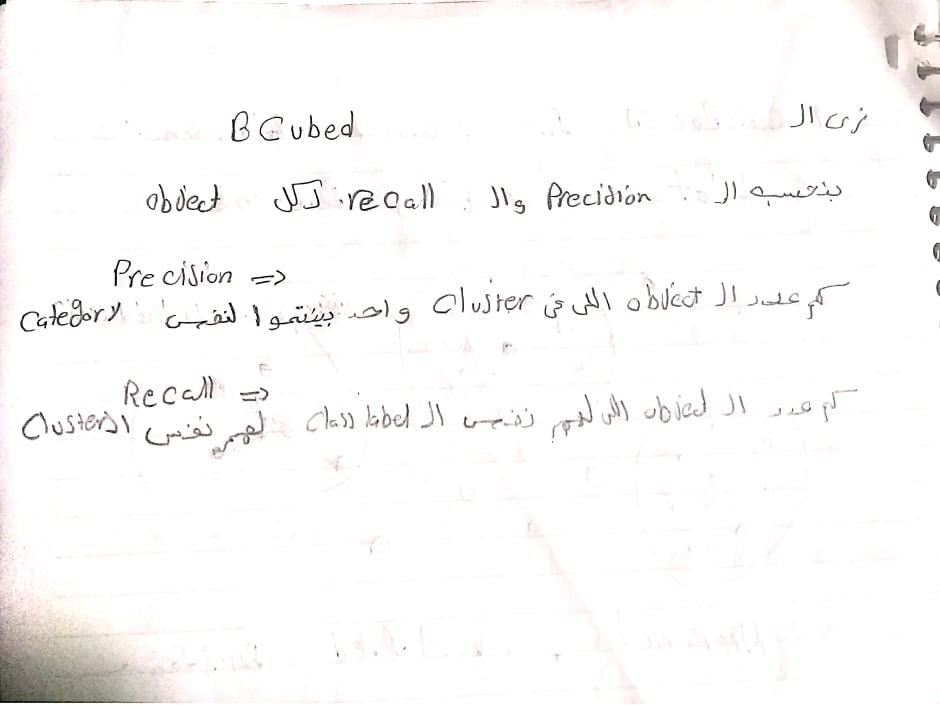


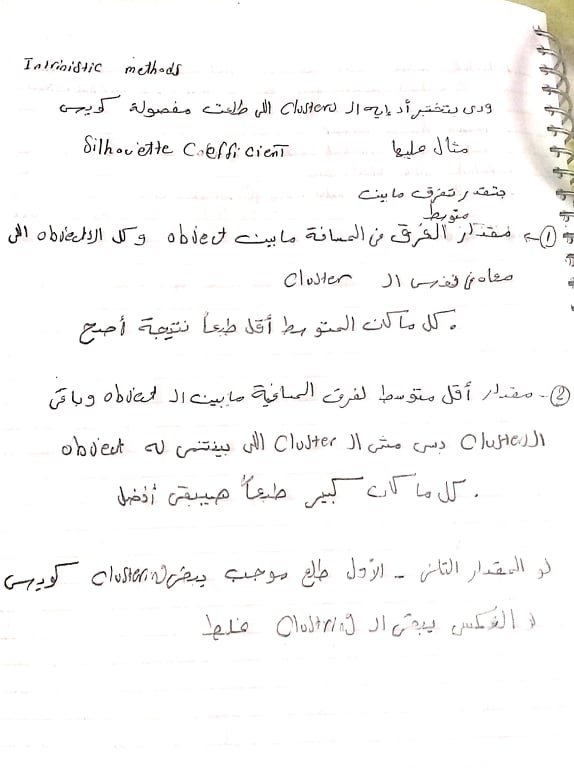


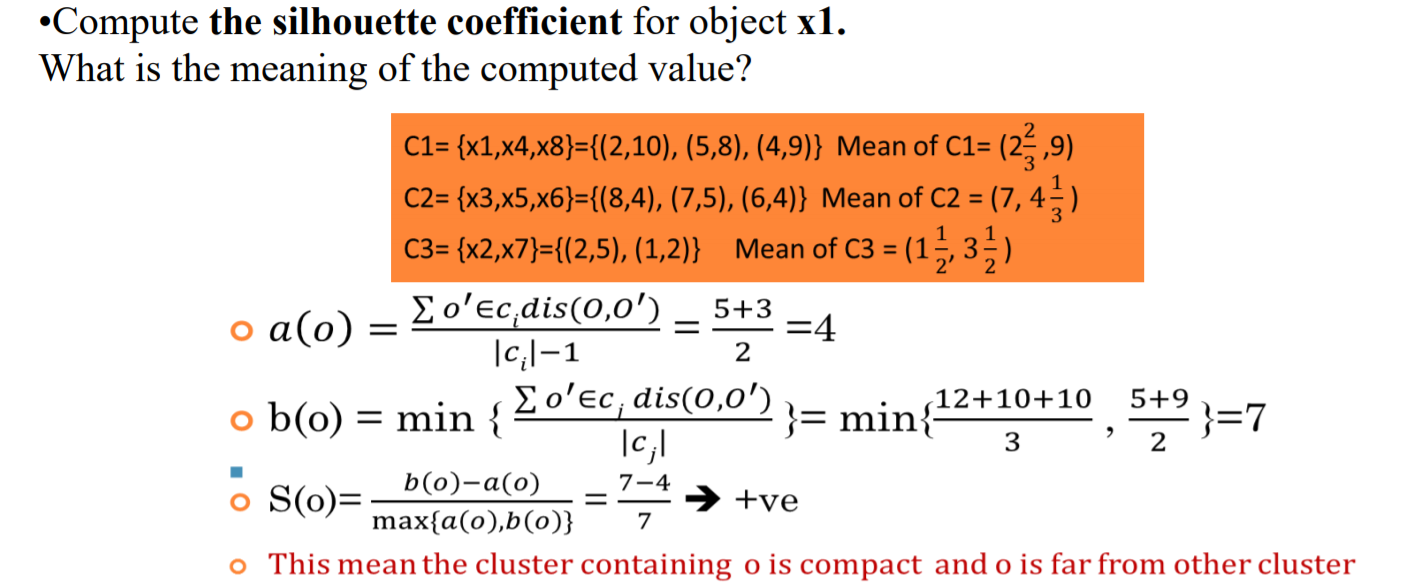
# EVALUATION OF CLUSTERING ASSESSING CLUSTERING TENDENCY











1. Define: A centroid in k-means.

Centroid: Is the mean of a cluster

2. Define: A core point in DBSCAN

3. Define: association and correlation analysis. Give an example

4. Define: cluster analysis. Give an example

5. Define: Data Cleaning, Data integration, Data reduction, Data transformation, Discretization 6. Define: outlier analysis. Give an example

7. Define: regression. Give an example

8. Give an example for nonparametric data reduction strategies.

9. Give an example for parametric data reduction strategies.

10. How does K-means differ from DBSCAN

11. How to assess the goodness of a rule?

12. How you can solve missing values problems

13. If a person’s height is measured in inches then what kind of attribute you will use? 14. If the correlation coefficient of the items bred and rice is equal to 1.5. This means what? 15. If the covariance of the items bred and rice is equal to 1. This means what? 16. If the information gain of age and income attributes are 0.24 and 0.024 respectively which one you will chose as the splitting attribute 17. If the lift measure of the items bred and rice is equal to 0.5. This means what? 18. If the lift measure of the items bred and rice is equal to 1. This means what? 19. If the lift measure of the items bred and rice is equal to 1.5. This means what? 20. If the mean is equal to the median then this might be an indication that the data is what?

21. If the mean is larger than the median then this might be an indication that the data is what? 22. If the mean is smaller than the median then this might be an indication that the data is what? 23. If you have 100 values in my data and I add 5.0 to all of the values, then how will this change the median? 24. If you have 100 values in my data and I add 5.0 to all of the values, then how will this change the median? 25. List the Cluster Analysis Methods 26. List the Major Preprocessing Tasks That Improve Quality of Data 27. List the steps of knowledge discovery 28. List the transformation strategies 29. List the types of outliers. Give an example for each one. 30. The confidence for the association rule {bread} → {milk, diapers} was determined to be 0.95. What does the value 0.95 mean? 31. The support for the association rule {bread} → {milk, diapers} was determined to be 0.95. What does the value 0.95 mean? 32. What are rules conflicts? How can you solve it? 33. What are the data smoothing techniques? 34. What are the different strategies of data reduction? 35. What are the main advantages and disadvantages of Decision Tree classification algorithms? 36. What are the terminating conditions in decision tree induction? 37. What classifiers are normally considered to be easy to interpret? 38. What clustering algorithms can find clusters of arbitrary shape? 39. What data mining task should be used to detect fraudulent usage of credit cards? 40. What is over fitting? Briefly describe one method to prevent over-fitting in classification trees. 41. What is the Apriori property? 42. What is the bootstrap sampling?

43. What is the different between noise and outlier? 44. What is the different between symmetric and asymmetric binary attribute? 45. What is the five numbers summery of the data? How is it represented graphically? 46. What is the majority voting? When you use it? 47. What is the means of association rule computer → webcam (60%, 100%) 48. What is the mode of the data? What is the mean of (bimodal, trimodel) 49. What is the problem that related to calculate the mean? How you can fix it? 50. What is the problem that related to use global constant to fill in the missing values. 51. What is the redundant attribute? How can you detect it? 52. What Kinds of Data Can Be Mined? 53. What Kinds of Patterns Can Be Mined? 54. When are objects q & m density-connected. 55. When is object p density-reachable from another object q? True or False 56. The silhouette coefficient is a method to determine the natural number of clusters for hierarchical algorithms density-based algorithms 57. All continuous variables are ratio 58. Association rules provide information in the form of "if-then" statements. 59. Attributes are sometimes called variables and objects are sometimes called observations 60. Binary variables are sometimes continuous 61. Cluster is the process of finding a model that describes and distinguishes data classes or concepts. 62. Computing the total sales of a company. Is a data mining task? 63. Correlation analysis divides data into groups that are meaningful, useful, or both.

64. Database mining refers to the process of deriving high-quality information from text. 65. Dissimilarity matrix stores n data objects that have p attributes as an n-byp matrix 66. Dividing the customers of a company according to their profitability. is a data mining task? 67. For an association rule, if we move one item from the right-hand-side to the left-hand-side of the rule, then the confidence will never change. 68. If all the proper subsets of an itemset are frequent, then the itemset itself must also be frequent. 69. In decision tree algorithms, attribute selection measures are used to rank attributes 70. In decision tree algorithms, attribute selection measures are used to reduce the dimensionality 71. In lazy learner we interest in the largest distance. 72. Intrinsic methods measure how well the clusters are separated 73. Multimedia Mining is the application of data mining techniques to discover patterns from the Web. 74. Regression is a method of integration 75. Strategies for data transformation include chi-square test 76. The Pruning make the decision tree more complex 77. An object is an outlier if its density is equal to the density of its neighbors. 78. A common weakness of association rule mining is that it is not produce enough interesting rules 79. Accuracy is interestingness measures for association rules 80. Binning is a method of reduction 81. Core object is an object whose -neighborhood contains objects less than MinPts 82. Correlation analysis is used to eliminate misleading rules. 83. Correlation is a method of cleaning

84. Data matrix stores a collection of proximities for all pairs of n objects as an n-by-n matrix 85. Extracting the frequencies of a sound wave. Is a data mining task? 86. Incomplete data problem can be solved by binning 87. K-Nearest Neighbor Classifiers do classification when new test data is available 88. Median is a value that occurs most frequently in the attribute values 89. Mode is a middle value in set of ordered values 90. One strength of a Bayesian Classifier is that it can be easily trained 91. Outlier analysis is a method of transformation 92. Predicting the outcomes of tossing a (fair) pair of dice. Is a data mining task? 93. Recall is interestingness measures for association rules 94. Redundancy is an important issue in data cleaning 95. Sampling methods smooth noisy data 96. Sorting a student database based on student identification numbers. Is a data mining task? 97. The bottleneck of the Apriori algorithm is caused by the number of association rules 98. The goal of clustering analysis is to maximize the number of clusters 99. the object is local outlier if it is deviate significantly from the rest of the dataset 100. The silhouette coefficient is a method to determine the natural number of clusters for hierarchical algorithms