PASS TASK (Task 4.1)

About this task

Step-1

At the completion of week 4 modules, you are required to compete a lesson review to indicate what you have learnt and how you learnt it by submitting evidence requested at the end of this file.

Step-2

Your tutor will then review your submission and will give you feedback. If your submission is incomplete the tutor will ask you to include missing parts. Tutor can also ask follow-up questions, either to clarify something that you have submitted or to assess your understanding of certain topics.

Feedback and submission deadlines

Feedback deadline: Friday 8 Aug (No submission before this date means no feedback!)
Submission deadline: Before creating and submitting portfolio.

Evidence of Learning

1. Submit a summary report (pdf format) in Ontrack (https://ontrack.deakin.edu.au)
   1.1. Summarise the main points that is covered in week 3 and 4.
   1.2. Provide summary of your reading list – external resources, websites, book chapters, code libraries, etc.
   1.3. Reflect on the knowledge that you have gained by reading contents of this week with respect to machine learning.
   1.4. Attempt the quiz given in weekly content (3.13 and 4.18) and add screenshot of your score (>85% is considered completion of this task) in this report.
2. Complete the problem solving task given below and submit your code file (.ipynb) separately to OnTrack (https://ontrack.deakin.edu.au).

Problem Solving:

1. Load the “Obesity” dataset. Remove the class label “Nobeyesdad” and encode the rest variables in an appropriate way.
2. Select the optimum k value using Silhouette Coefficient and plot the optimum k values, is the optimum k the same as the number of classes? Explain why it is different or not.
3. Create clusters using Kmeans and Kmeans++ algorithms with optimal k value found in the previous problem. Report performances using appropriate evaluation metrics. Compare the results.
4. Now repeat clustering using Kmeans for 50 times and report the average performance. Again compare the results that you have obtained in Q3 using Kmeans++ and explain the difference.
5. Apply DBSCAN on this same Obesity dataset and find the optimum "eps" and "min_samples" value. Is the number of cluster same as the cluster found in Q2? Explain the similarity or differences that you have found between two solutions.
6. Load the “gene expression” dataset. Apply PCA on the genes for generating 3 principal components. Plot the first three components of the PCA.
7. Continue from question 6, what is the variance (%) covered by the first three components? Explain how is this percentage of variance computed?
8. Continue from question 6, apply KMeans on the original features of the gene dataset and the first three components returned by PCA. Compare the results using the given labels.