Homework 12 - Clustering

Unsupervised Learning - Clustering algorithms (60 points)

Instruction:

- This is a group-work assignment!
- You are expected to submit the **.ipynb** file and the exported **.html**.
- Only one member in each group needs to submit the assignment. It will be automatically submitted for the rest of group members.

  Good luck and enjoy machine learning!

Data: The Wine data set!

In this exercise, I want you to work with Wine data set. The goal is to find the optimal number of customer segments based on given features. (Note that the data set already comes with a variable named customer_Segment which has 3 segments (clusters). Let's see what's the reasoning behind that 3 segments.

Question 2 Hierarchical Clustering (30 points)

Use the **AgglomerativeClustering()** class from sklearn.cluster module and answer the following questions:

1. Explain why it is not a good idea to use dendrogram to find the optimal number of clusters in this exercise!
2. Fit the hierarchical clustering to the data set and compare clusters with customer_segments. Use n_clusters=3 even if 3 is not the optimal number of clusters
3. Find the accuracy of your clustering method. (**important note**: we know that Kmean clustering is an unsupervised model so finding the accuracy of the model does not make sense. However, in this example we can use the customer_Segment variable as target and construct the confusion matrix based on that. Finally you can report the accuracy for this set up)

Question 1 K-Mean Clustering (30 points)

Use the **KMeans()** class from sklearn.cluster module and answer the following questions:

1. Find the optimal number of clusters using the elbow method (clusters vs WCV)
2. Fit the K-Means to the data set and compare clusters with customer_segments. Use n_clusters=3 even if 3 is not the optimal number of clusters. We want to figure out what was the logic behind segmenting the customers into 3 groups in the original data set.
3. Report the accuracy of your clustering method. (again use the customer_Segment variable as target and construct the confusion matrix based on that)

Good luck and enjoy machine learning!