Task 3

Model the data by treating it as **either** a *Classification* or *Clustering* Task, depending on your choice.

You must use **two** different models (i.e. two Classification models, or two Clustering models), and when building each model, it must include the following steps:

- Select the appropriate features
- Select the appropriate model (e.g. *DecisionTree* for classification) from *sklearn*.
- If you choose to do a *Classification* Task,
  - Train and evaluate the model appropriately.
  - Train the model by selecting the appropriate values for each parameter in the model. You need to show how you choose this values, and justify why you choose it.
- If you choose to do a *Clustering* Task,
  - Train the model by selecting appropriate values for each parameter in the model.
    - *Show* how do you choose this value, and *justify* why you choose it (for example, *k* in the *k*-means model).
  - *Determine the optimal number of clusters, and justify* – Evaluate the performance of the clustering model by:
    - Checking the clustering results against the true observation labels
    - Constructing a "confusion matrix" to analyse the meaning of each cluster by looking at the majority of observations in the cluster. (You can do this by using a pen and a piece of paper, as we did in Practical Exercise; if you prefer, you can also explore how to do this step directly in IPython.)

After you have built two Classification models, or two Clustering models, on your data, the next step is to **compare** the models. You need to include the results of this comparison, including a recommendation of which model should be used, in your report (see next section).