Fourth Industrial Revolution (4IR) Summer School

Data **A**nalysis **F**oundations – **D**ay **1 (A)** exercises

---

Question 1 [Statistical Analysis]

Examine the supplemented csv file named (ExperimentData.csv), you will find the error rate scores for two groups (A and B). We formulated the following hypothesis:

H_0: There is no difference in error rate means between Group A and Group B

where the alternative

H_a: There is a difference in error rate means between Group A and Group B

Write a Python program to perform a statistical analysis to test the above hypothesis and explain the results of your analysis. You need to check if given data is normally distributed, and then perform the appropriate statistical test.

Question 2 [Correlation]

Consider the following python code:

```
data1 = np.array([44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 46, 47, 48, 60.1])
data2 = np.array([2.6, 3.1, 2.5, 5.0, 3.6, 4.0, 5.2, 2.8, 4, 4.1, 4.5, 3.8])
```

- Study the correlation between the two variables data1 and data2, and show your findings.

- Study again the correlation between these two randomly generated data

```
from numpy.random import rand
from numpy.random import seed
from matplotlib import pyplot
# seed random number generator
seed(1)
# prepare data
data1 = rand(1000) * 20
data2 = data1 + (rand(1000) * 10)
```

Question 3 [Statistical Analysis]

Examine the supplemented csv file named (ClassifiersAccuracy.csv), you will find the accuracy scores for three machine learning classifiers. We formulated the following hypothesis:

H_0: There is no difference in accuracy means between DT, SVM and ANN

where the alternative

H_a: There is a difference in accuracy means between DT, SVM and ANN

Write a Python program to perform a statistical analysis to test the above hypothesis and explain the results of your analysis. You need to check if given data is normally distributed, and then perform the appropriate statistical test.