# Machine Learning: Practical Work

Mourad TERZI

## Unsupervised Learning

### Description of the dataset

**Sources:**

- https://husson.github.io/data.html
- http://factominer.free.fr/livre/

**Dataset Overview**:

- Contains temperature data for **35 European cities**.
- Cities are divided into:
  - **Capitals**: Rows 1 to 24 (e.g., Amsterdam to Stockholm).
  - **Major cities (non-capitals)**: Rows 25 to 35 (e.g., Anvers to Zurich).

**Variables**:

- **Quantitative**: 16 variables.
  - Indexed from 1 (Janvier) to 12 (Décembre) for monthly temperatures.
  - Variables 13 (Moyenne) to 16 (Longitude) are supplementary.
- **Qualitative**: 1 variable (Région), indicating the geographic location of the city.

**Scope of the Exercise**:

- Focuses only on **European capitals** (23 rows).
- Analyzes variables **1 to 12** (monthly temperatures).

**Filtered Dataset for Study**:

- Rows: **23** (European capitals only).
- Columns: **12** (Monthly temperatures).

### Part 1: Principal Component Analysis (PCA)

1. Is it necessary to center and scale the data before applying PCA? Justify your answer.
2. Perform Principal Component Analysis (PCA) on the given dataset.
3. What does the variance of each component represent? What is the total variance explained by all the components?
4. Display the individuals (data points) on the first two principal components (PCA axes). Analyze the resulting graph and describe any patterns or clusters.

### Part 2: K-Means Clustering

1. What is the difference between the Silhouette criterion and the Elbow method for selecting the best value of K for K-means? Explain both methods.
2. Use the Silhouette method to determine the optimal value of K for K-means clustering.
3. Use the Elbow method to confirm the value of K obtained from the Silhouette method.

4. Apply K-means clustering using the K value selected through the Silhouette method.
5. Plot the individuals on the first two principal components (from the PCA) and visualize how the K-means clusters are distributed.
6. Perform a chi-square test between the ' Région' column in the dataset and the K-means labels. What do you observe from the results of the test?
7. Analyze the clusters obtained based on the following features:
    - Région
    - Latitude
    - Longitude
    - Altitude

### Part 3: DBSCAN Clustering

1. Select the optimal value of 'eps' for DBSCAN using the K-nearest neighbors' method with K=3. Explain the process and plot the distance graph for K-nearest neighbors.
2. Apply DBSCAN with the optimal value of 'eps' obtained in the previous step and analyze the obtained clusters.

### Part 4: DBSCAN with Reduced Data

1. Remove the cities Athens, Madrid, Rome, and Lisbon from the dataset.
2. Perform PCA on the reduced dataset. Display the individuals on the first two principal components and analyze the resulting graph.
3. Apply the K-nearest neighbors' method with K=4 to select the optimal value of 'eps' for DBSCAN. Explain the process and show the graph for K-nearest neighbors.
4. Apply DBSCAN to the reduced dataset using the optimal 'eps' value and analyze the resulting clusters.

## Supervised Learning

### Description of the dataset

**Source:**

- https://husson.github.io/data.html

**Dataset Overview**:

- Decathlon dataset with 41 athletes (rows).
- In the original file, athletes are classified into 4 classes, but for this exercise, we are working with 3 labeled classes: 1, 2 and 3.
- One qualitative column "Classe" containing the label for each athlete.

**Variables**:

- 10 quantitative variables (athletes' performances).
- 2 quantitative variables (Classement and Points).
- One qualitative variable: competition (2004 Olympic Games or Decastar).

**Scope of the Exercice:**

- Focuses only on classes 1, 2 and 3.
- Analyzes 10 athletes' performances variables.

**Filtered Dataset for study:**

- Rows: 38 (classes 1, 2 and 3).
- Columns: 10 (athletes' performances).

## Part 1: KNN

1. Is the cross-validation important for training KNN or another machine learning model on our dataset? Justify your answer.
2. Train the KNN model with different values of k (e.g., k=3, k=5, k=7).
    - Use cross-validation to select the best k based on performance metrics like accuracy and F1-score.

1. Analyze the confusion matrix of KNN with the optimal value of k returned by cross-validation.
2. Display a classification report for detailed performance insights.
3. Plot the decision boundaries generated by KNN using a 2D graph (example, graph of PCA).

## Part 2: SVM

1. Use cross-validation to select the best combination of C and kernel of SVM based on performance metrics such as accuracy and F1-score.
    1. C: 0.1, 1, 10, 100.
    2. kernel: [linear, rbf, sigmoid]
2. Analyze the confusion matrix of SVM with the optimal value of C and kernel returned by cross-validation.
3. Display a classification report for detailed performance insights.
4. Plot the decision boundaries generated by SVM using a 2D graph (example, graph of PCA).

## Part 3: Models comparison

1. Compare the performance of the best KNN model (with the optimal k) and the best SVM model (with the optimal C and kernel).

## Part 4: To go further

1. Explore and propose another model of your choice, such as decision tree or neural network.
2. Compare the performance of the proposed model to the best KNN and best SVM from the first two parts.
3. Perform feature selection to identify the most relevant features in the dataset and use the selected features to train one model of your choice: either KNN or SVM.