### ****Unsupervised Clustering****

In this exercise, we applied an unsupervised clustering algorithm (K-means) on a corpus of text messages (spam.csv), with the goal of exploring the possibility of automatically distinguishing spam messages from ham messages (non-spam), without using the labels.

#### ****1. Data Preparation****

* **TF-IDF Vectorization**: The text messages were transformed into numerical vectors using the TF-IDF method, keeping a maximum of 1000 distinct words, and removing English stop words.
* **Dimensionality Reduction**: A **Principal Component Analysis (PCA)** was applied to project the vectorized data into a 2-dimensional space, to allow visualization.
* **Clustering**: The **K-means algorithm** was then applied with n\_clusters = 2 to group the data without knowing the actual labels.

#### ****2. Visual Results****

* The **left plot** represents the clusters generated by K-means: the points are colored according to the group each message was assigned to (cluster 0 or 1).
* The **right plot** represents the real classes: 0 for ham (non-spam), 1 for spam.

#### ****3. Interpretation****

* **Visually**, K-means succeeds in separating the data into two distinct groups, with a certain coherence compared to reality.
* By comparing the two plots:
  + We observe that the majority of points identified as spam in the real labels (right plot, orange points) correspond to a well-separated cluster in the left plot.
  + The ham messages, which are much more numerous, are also grouped in the other cluster.
* This performance is notable since the algorithm never had access to the real labels: it is therefore a completely unsupervised grouping.