


<b>Module : Analyse et fouille de données</b>	
<b>Responsable du Cours:</b> Bouaziz Souhir, Abbas Amal <b>Enseignants TP:</b> Barhoumi Chawki, Rekik Amal, Njeh Maissa	<b>Auditoire:</b> D-LSI-ADBD <b>A-U:</b> 2023-2024
<b>TP1 : Introduction au processus ECD</b> <i>Phases de prétraitement et de transformation</i>	

### **Introduction :**

Face à l'explosion continue du volume des données, le processus d'extraction des connaissances à partir des données **ECD** est devenu de plus en plus une nécessité dans divers domaines d'application. Ainsi, cette démarche s'avère cruciale dans des secteurs aussi variés que la recherche scientifique, la prise de décision en entreprise, la médecine, et bien d'autres.

Ce processus se déclenche par une phase de prétraitement des données pour avoir des informations, ces données sont ensuite analysées et traitées pour extraire des nouvelles représentations appelée connaissances.

### **Objectives :**

Après avoir maîtriser les outils de manipulation des dataframe dans le TP0, l'objectif de ce TP est de maîtriser les bibliothèques **pandas**, **numpy**, **cv2** de python permettant la lecture, le nettoyage, et la transformation des données brutes.

### **Exercice1 :**

Soit le DataFrame suivant qui représente des informations sur des employés au sein d'une entreprise

	Nom	Date_emb	poste	Salaire	nb heures trav
0	Alice	30-30-3030	manger	NaN	18
1	Bob	15-02-2020	Ingenieur	90000.0	42
2	Charlie	01-06-2024	Developpeur	NaN	30
3	David	12-08-2020	Ingenieur	65000.0	38
4	Emma	16-05-2022	Ingenieur	83000.0	35
5	Alice	06-12-2023	manger	8000.0	18
6	Bob	15-02-2020	Ingenieur	90000.0	42
7	marwa	01-12-2013	Developpeur	73000.0	48

- 1) Télécharger dans votre notebook la base de données '*empl.csv*'
- 2) Interpréter le jeu de données
- 3) Vérifier l'intégrité du domaine des valeurs de la variable '*Date\_embau*'
- 4) Vérifier s'il existe des doublons dans votre base, si c'est le cas procéder à leur suppression.
- 5) Remplacer toutes les valeurs NaN par Zéro
- 6) Déduire si le traitement de la question précédente a conduit à l'apparition de valeurs aberrantes dans la base de données, si c'est le cas remplacer cette valeur par la valeur max de la variable.

- 7) Appliquer le codage nécessaire pour transformer les valeurs de la variable Poste en des valeurs numériques. (Exemple ; manger :0, Ingénieur :1, Développeur : 2)
- 8) Calculer la matrice Y des données centrées et la matrice Z de données centrées et réduites.
- 9) Calculer la matrice RX des corrélations de notre matrice de données et la matrice VZ des variances et covariances de Z. Commenter.

## Exercice 2 : ‘prétraitement des images’

On souhaite récupérer et représenter la base des images ‘*images\_data*’ existant dans votre Classroom.

- 1) Ecrire une fonction qui permet de lire ces images et les stocker dans une liste.
- 2) Vérifier si toutes les images ont la même dimension. Si ce n'est pas le cas, changer leur taille pour qu'elles aient toutes la même dimension (64,64).
- 3) Créer un dataframe comprenant ces images, chaque image doit être transformée en un vecteur ligne de dimensions 64 x 64, les colonnes correspondent aux pixels.
- 4) Déterminer le maximum et le minimum des niveaux de gris dans le dataframe, déduire l'intégrité des valeurs des niveaux de gris des pixels.
- 5) Ajouter deux colonnes au dataframe nommées étendu et médian représentant l'étendu et la moyenne de chaque observation.
- 6) Déterminer puis afficher les deux images les plus similaires en utilisant ces deux propriétés. Interpréter les résultats obtenus.